IDIAP RESEARCH REPORT

HMM Mixtures (HMM2) for Robust Speech Recognition

Katrin Weber[a]

IDIAP RR 03-34

June 2003

PUBLISHED AS

Docteur ès Sciences thesis no. 2790 (2003),
Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland.

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] with the Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), CH-1920 Martigny, Switzerand, and with the Swiss Federal Institute of Technology Lausanne (EPFL), CH-1015 Lausanne, Switzerland

To Peter

for his love

and his continuous support

in good and bad times throughout this thesis

To Laura Lou

for her smiles

and the energy they gave me

when I needed it most

To my parents

for their perspective

about the relative importance of a thesis

and other things in life

Abstract

State-of-the-art automatic speech recognition (ASR) techniques are typically based on hidden Markov models (HMMs) for the modeling of temporal sequences of feature vectors extracted from the speech signal. At the level of each HMM state, Gaussian mixture models (GMMs) or artificial neural networks (ANNs) are commonly used in order to model the state emission probabilities. However, both GMMs and ANNs are rather rigid, as they are incapable of adapting to variations inherent in the speech signal, such as inter- and intra-speaker variations. Moreover, performance degradations of these systems are severe in the case of unmatched conditions such as in the presence of environmental noise. A lot of research effort is currently being devoted to overcoming these problems.

The principal objective of this thesis is to explore new approaches towards a more robust and adaptive modeling of speech. In this context, different aspects of the modeling of speech data with HMMs and GMMs are investigated. Particular attention is given to the modeling of correlation. While correlation between different feature vectors (corresponding to temporal correlation) is typically modeled by the HMM, correlation between feature vector components (e.g., correlation in frequency) is modeled by the GMM part of the model. This thesis starts with the investigation of two potential ways to improve the modeling of correlation, consisting of (1) a shift of the modeling of temporal correlation towards GMMs, and (2) the modeling of correlation within each feature vector by a particular type of HMM. This leads to the development of a novel approach, referred to as "HMM2", which is a major focus of this thesis.

HMM2 is a particular mixture of hidden Markov models, where state emission probabilities of the temporal (primary) HMM are modeled through (secondary) state-dependent frequency-based HMMs. Low-dimensional GMMs are used for modeling the state emission probabilities of the secondary HMM states. Therefore, HMM2 can be seen as a generalization of conventional HMMs, which they include as a particular case. HMM2 may have several advantages as compared to standard systems. While the primary HMM performs time warping and time integration, the secondary HMM performs warping and integration along the frequency dimension of the speech signal. Frequency correlation is modeled through the secondary HMM topology. Due to the implicit, non-linear, state-dependent spectral warping performed by the secondary HMM, HMM2 may be viewed as a dynamic extension of the multi-band approach. Moreover, this frequency warping property may result in a better, more flexible modeling and parameter sharing. After an investigation of theoretical and practical aspects of HMM2, encouraging recognition results for the case of speech degraded by additive noise are given.

Due to the spectral warping property of HMM2, this model is able to extract pertinent structural information of the speech signal, which is reflected in the trained model parameters. Consequently, such an HMM2 system can also be used to explicitly extract structures of a speech signal, which can then be converted into a new kind of ASR features, referred to as "HMM2 features". In fact, frequency bands

with similar characteristics are supposed to be emitted by the same secondary HMM state. The warping along the frequency dimension of speech thus results in an adaptable, data-driven frequency segmentation. In fact, as it can be assumed that different secondary HMM states model spectral regions characterized by high and low energies respectively, this segmentation may be related to formant structures. The application of HMM2 as a feature extractor is investigated, and it is shown that a system combining HMM2 features with conventional noise-robust features yields an improved speech recognition robustness. Moreover, a comparison of HMM2 features with formant tracks shows a comparable performance on a vowel classification task.

The structure of this thesis is as follows. After an introduction of the state-of-the-art in automatic speech recognition, the shifting of the modeling of time and frequency correlation towards GMMs and HMMs respectively is briefly investigated. Then, the HMM2 approach is introduced, and its theory is presented. This is followed by an experimental evaluation of HMM2 on a speech recognition task. The application of HMM2 as feature extractor is investigated, and HMM2 features are compared to formants. Finally, the most important results are summarized, and possible future research directions are outlined.

Version abrégée

Les systèmes de l'état de l'art de reconnaissance automatique de la parole sont typiquement basés sur des modèles de Markov cachés (Hidden Markov Models, HMMs) qui modélisent des séquences temporelles de vecteurs acoustiques extraits du signal de parole. Au niveau de chaque état du HMM, des mixtures de Gaussiennes (Gaussian Mixture Models, GMMs) ou des réseaux de neurones artificiels (Artificial Neural Networks, ANN) sont le plus souvent employés pour la modélisation des probabilités d'émissions. Cependant, les GMM et les ANN sont assez rigides, n'étant pas capables de s'adapter aux variations inhérentes du signal de parole, telles que les variations inter- et intra-locuteur. Beaucoup d'effort de recherche est actuellement mis en oeuvre afin de proposer des solutions à ces problèmes.

L'objectif principal de cette thèse est d'explorer de nouvelles approches vers une modélisation plus robuste et adaptative du signal de parole. Dans ce contexte, des aspects différents de la modélisation des données représentant la parole par des HMMs et des GMMs sont étudiés. Alors que la corrélation entre les différents vecteurs acoustiques (correspondant à la corrélation temporelle) est typiquement modélisée par le HMM, la corrélation entre les coefficients des vecteurs acoustiques (par exemple, la corrélation en fréquence) est modélisée par le GMM. Cette thèse commence avec une étude de deux possibilités pour améliorer la modélisation de la corrélation : (1) un décalage de la modélisation de la corrélation temporelle vers les GMMs, et (2) la modélisation de la corrélation entre les composants de chaque vecteur acoustique avec un type particulier de HMM. Cela mène au développement d'une nouvelle approche, appelée "HMM2", qui constitue un des focus principaux de cette thèse.

Un HMM2 est une mixture particulière de modèles de Markov cachés, où les probabilités d'émission de chaque état du HMM temporel (dit primaire) sont modélisées avec des HMMs (dit secondaires), travaillant dans le domaine des fréquences, et qui dépendent de l'état du HMM primaire. Des GMMs de basse dimension sont utilisés pour la modélisation des probabilités d'émission de chaque état du HMM secondaire. Par conséquent, l'approche HMM2 peut être vue comme une généralisation des HMMs conventionnels, qui constituent en fait un cas particulier des HMM2. Un HMM2 peut avoir de nombreux avantages par rapport aux systèmes standards. Tandis que le HMM primaire effectue un "warping" (c.a.d. un regroupement) et une intégration dans la dimension temporelle, le HMM secondaire effectue un warping et une intégration dans la dimension fréquentielle du signal de parole. La corrélation en fréquence est modélisée par la topologie du HMM secondaire. En raison du warping implicite et non-linéaire effectué par le HMM secondaire, un HMM2 peut être vu comme une extension de l'approche multi-bande. En outre, le warping en fréquence peut résulter en une meilleure modélisation, plus flexible, permettant en plus un partage efficace des paramètres. Après une étude des aspects théoriques et pratiques de l'approche HMM2, des résultats encourageants pour le cas de la reconnaissance de la parole bruitée additivement sont donnés.

Grace au warping du spectre effectué par le HMM2, ce modèle peut extraire des informations pertinentes sur la structure du signal de parole, ce qui est reflété dans les paramètres d'un modèle entraîné. Par conséquent, un tel HMM2 peut être employé afin d'extraire explicitement des structures d'un signal de parole. Ces structures peuvent être converties dans un nouveau type de coefficients, dit "features HMM2". En fait, des bandes de fréquences montrant une caractéristique similaire sont supposées être émises par le même état du HMM secondaire. Le warping dans la dimension des fréquences génère donc une segmentation adaptable en fonction des données. Comme on peut supposer que les états différents du HMM secondaire modélisent des régions de basses ou hautes énergies respectivement, cette segmentation peut être en relation avec les formants. L'applications du HMM2 comme extracteur de coefficients est étudié, et il est montré qu'un système qui combine ces "features HMM2" avec des coefficients conventionnels et robustes aux bruits obtient une amélioration de la robustesse en reconnaissance de la parole. De plus, une comparaison des "features HMM2" avec les traces de formants montre des résultats comparables pour la tache de la classification de différentes voyelles.

La thèse est structurée ainsi : Après une introduction de l'état de l'art en reconnaissance automatique de la parole, le décalage de la modélisation de la corrélation temporelle et fréquentielle vers les GMMs et vers les HMMs respectivement est étudié. Ensuite, l'approche HMM2 est introduite, en commençant par la théorie. Ceci est suivi par une évaluation des HMM2 pour la reconnaissance de la parole. L'application des HMM2 comme extracteur de coefficients est étudiée, et les "features HMM2" sont comparés aux formants. Finalement, les résultats les plus importants sont récapitulés, et des directions possibles pour la recherche future sont données.

Kurzfassung

Algorithmen zur automatischen Spracherkennung, die dem aktuellen Stand der Technik entsprechen, basieren in der Regel auf Hidden-Markov-Modellen (Hidden Markov Models, HMMs), die die zeitliche Abfolge von Merkmalsvektoren beschreiben. Die Emissions-Wahrscheinlichkeiten werden für jeden einzelnen Zustand des Modells meist als Mischung von Gaußkurven (Gaussian Mixture Models, GMMs) oder als künstliche neuronale Netze (Artificial Neural Networks, ANNs) modelliert. Jedoch sind sowohl GMMs als auch ANNs relativ unflexibel und können sich nicht an die für Sprachsignale typischen Variationen (z.B. zwischen verschiedenen Sprechern oder zwischen verschiedenen Aussprachevarianten desselben Sprechers) anpassen. Zudem versagen sie häufig unter gegenüber dem Trainingsfall veränderten Bedingungen, z.B. bei Hintergrundgeräuschen. Zur Zeit wird intensiv an Lösungen zu diesen Problemen geforscht.

Das hauptsächliche Anliegen dieser Arbeit ist es, neue Ansätze für eine robustere und adaptive Sprachmodellierung zu erforschen. In diesem Zusammenhang werden verschiedene Aspekte der Modellierung des Sprachsignals mittels HMMs und GMMs untersucht. Besondere Aufmerksamkeit wird der Modellierung von Korrelationen geschenkt. Während die Korrelation zwischen verschiedenen Merkmalsvektoren (zeitliche Korrelation) typischerweise mit einem HMM beschrieben wird, so ist das GMM für die Modellierung der Korrelation zwischen den einzelnen Komponenten eines Merkmalsvektors (Korrelation bezüglich der Frequenz) verantwortlich. Diese Arbeit beginnt mit einer Untersuchung von zwei möglichen Wegen, die Modellierung der Korrelation zu verbessern. Zum einen wird eine Verschiebung der Modellierung zeitlicher Korrelation in Richtung eines GMM untersucht. Zum anderen wird die Modellierung der Korrelation zwischen den Komponenten eines Merkmalsvektors mit einem speziellen HMM erforscht. Dies führt zur Entwicklung eines neuen Ansatzes, der als "HMM2" bezeichnet wird und der den Fokus dieser Arbeit bildet.

HMM2 ist eine besondere Mischung aus HMMs, bei der die Emissions-Wahrscheinlichkeiten des zeitlichen (primären) HMM durch zustandsabhängige, frequenzbasierte (sekundäre) HMMs beschrieben werden. GMMs niedriger Dimension werden für die Modellierung der Emissions-Wahrscheinlichkeiten der Zustände des sekundären HMM genutzt. Deshalb können konventionelle HMMs als Spezialfall von HMM2 betrachtet werden. Verglichen mit Standard-HMMs hat HMM2 verschiedene potentielle Vorteile. Während das primäre HMM ein Warping (d.h. ein Verziehen) und eine Integration über die zeitliche Dimension ausführt, vollzieht das sekundäre HMM ein Warping und eine Integration über die Frequenzen des Sprachsignals. Korrelationen über der Frequenz werden durch die Topologie des sekundären HMM beschrieben. Wegen des impliziten, nicht-linearen, zustands-abhängigen spektralen Warpings des sekundären HMM kann HMM2 als eine dynamische Erweiterung des "Multi-Band-Ansatzes" betrachtet werden. Außerdem kann dieses Frequenz-Warping zu einer besseren, flexibleren Modellierung und zu einer gemeinsamen Parameter-Nutzung führen. Nach einer Untersuchung von the-

oretischen und praktischen Aspekten von HMM2 werden Erfolg versprechende Resultate für die Erkennung von additiv verrauschter Sprache präsentiert.

Als eine Folge des spektralen Warpings extrahiert HMM2 automatisch relevante strukturelle Informationen der Sprache, welche in den trainierten Parametern widergespiegelt werden. Demzufolge kann HMM2 auch zur Extraktion expliziter Strukturen aus einem gegebenen Sprachsignal eingesetzt werden. Diese können dann in eine neue Art von Merkmalsvektoren umgewandelt werden, welche "HMM2-Merkmalsvektoren" genannt werden. Tatsächlich ist anzunehmen, dass Frequenz-Bänder mit ähnlicher Charakteristik vom gleichen Zustand des sekundären HMM emittiert werden. Deswegen führt das Frequenz-Warping zu einer anpassungsfähigen, datengesteuerten Segmentierung des Sprachsignals entlang der Frequenz-Axe. Da angenommen werden kann, dass Regionen hoher bzw. niedriger spektraler Energie durch unterschiedliche Zustände des sekundären HMM beschrieben werden, könnte sich diese Segmentierung an den Formanten des Sprachsignals orientieren. Die Anwendung von HMM2 zur Extraktion von HMM2-Merkmalsvektoren wird untersucht und es wird gezeigt, dass die Kombination von konventionellen (gegenüber Rauschen robusten) Merkmalsvektoren und von HMM2-Merkmalsvektoren zu einer verbesserten Robustheit der Spracherkennung führt. Außerdem zeigt ein Vergleich zwischen HMM2-Merkmalsvektoren und Formantverläufen eine vergleichbare Leistung bei der Klassifikation von Vokalen.

Diese Arbeit ist wie folgt strukturiert: Nach einer Einführung in die automatische Spracherkennung wird die angesprochene Verschiebung der Modellierung von Zeit- und Frequenz-Korrelation in Richtung GMM und HMM untersucht. Anschießend wird der HMM2-Ansatz und die ihm zugrunde liegende Theorie präsentiert. Es folgt eine experimentelle Bewertung des Ansatzes mittels einer Spracherkennungs-Aufgabe. Die Anwendung von HMM2 zur Gewinnung von HMM2-Merkmalsvektoren wird untersucht und die so extrahierten HMM2-Merkmalsvektoren werden mit den Formanten eines Sprachsignals verglichen. Abschließend werden die wichtigsten Resultate der Arbeit zusammengefasst und es werden mögliche Richtungen für die zukünftige Forschung aufgezeigt.

Acknowledgements

This thesis was hosted by IDIAP, Martigny, and carried out at EPFL, Lausanne, Switzerland. I feel very privileged for having been given the opportunity to work and study in such an extraordinary environment, in different respects. I would like to express my deepest thanks to all those who directly or indirectly helped making this thesis possible, who provided all the means for this thesis ranging from scientific advice to moral support, and, last but not least, the necessary financing. This thesis has been supported by the Swiss National Science Foundation with grants Nr. FN 2100-50742.97/1 and Nr. FN 2000-059169.99/1, by the NCCR (IM)2 project on Interactive Multimodal Information Management, and by additional funding from IDIAP.

I would like to thank Prof. Hervé Bourlard for directing this thesis throughout the years, in spite of the fact that it was not always easy. I also profited from the scientific advice of different advisors, and from discussion with many of my colleagues, for which I am most grateful. In particular, I would like to thank Dr. Chafic Mokbel and Dr. Souheil Ben-Yacoub for having supported my early work. I am also especially indebted to Dr. Samy Bengio, whose competence, sincerity, and integrity greatly contributed to the quality of work and scientific exchange. His advice was always highly appreciated. Special thanks go to Febe de Wet and her colleagues from the University of Nijmegen for our fruitful collaboration. This document also profited from valuable comments, corrections, and other help from my colleagues and friends, particularly from Dr. Iain McCowan, Darren Moore, Jo Moore, Shajith Ikbal, Dr. Sebastian Möller, and Gianni Pante. I would also like to thank the president and the members of my thesis committee for having accepted to judge this work. Many thanks to the IDIAP system group for their support with hard- and software, and to the IDIAP administration and secretariat for their help with organizational problems.

The years I spent working on my thesis coincided with a particularly turbulent time in my life. It was a time of interesting professional, but also of profound personal experiences. The time of my thesis was a period of intensive and constant learning, both about speech recognition research and about some more general aspects of the functioning of the world. I appreciate all the lessons I had the chance to learn. I would like to express my gratitude to my family and friends who helped me navigate through these ups and downs. More particularly, I would like to thank those who provided some practical help at difficult times, but also those who simply shared some good moments with me over the years, mountaineering, hiking, mountain biking, snowboarding, dancing, international cooking, etc. All this helped me to get my mind off work, and, full of new energy, back on track again.

However, those who gave the most in order for this thesis to come to an end are my husband Peter and our daughter Laura Lou. While Laura Lou had to put up with her mummy going to work from a very tender age, and with long and not always easy days in the nursery, Peter greatly compensated for this with his presence, whenever possible. In a difficult personal situation, he encouraged me very much in the decision to pursue my work, knowing that this meant an extraordinary additional load for him also. In addition to him being such a marvellous father and husband, he gave me a lot of practical as well as emotional support. This thesis would simply not have been possible without him.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the most fundamental characteristics distinguishing humans from all other living beings is the use of high level spoken language for communication. The importance of the use of speech is already reflected in old science fiction movies, where computers understood humans. For example, already in the 1960s, the makers of "2001: A Space Odyssey" depicted a sophisticated system that allowed humans to talk to computers. Also researchers were realizing that using spoken language is indeed one of the most natural and efficient ways for humans to communicate and that it could be very valuable to use such a communication scheme with computers as well. Today, as the scientific progress (as well as the industrial and economic environment) allow advanced technology for computer access, this concept is penetrating many fields of our daily lives. However, the use of spoken language for human-computer interaction is still restricted to specific, limited domains. For instance, current automatic speech recognition (ASR) systems typically only deal either with a limited vocabulary (such as in voice dialing applications which come with many commercial cellular phones), or they need to be adapted to a certain speaker (such as in dictation systems). Moreover, these ASR systems almost inevitably fail in spontaneous speech and in other difficult conditions such as under environmental noise, where human speech recognition hardly degrades.

A lot of research effort is currently spent in laboratories all around the world in order to alleviate these problems, and several approaches towards this goal are also investigated in our institution. The goal of improving ASR robustness defines also the framework of this thesis. More particularly, on the base of state-of-the-art speech recognition technology, new approaches towards a more robust and adaptive modeling of speech are investigated. To first give an introduction of the state-of-the-art in automatic speech recognition (ASR), this chapter starts with an illustration of the structure of a typical ASR system and a discussion of its major modules. Then, weaknesses of this standard system are reviewed, and some approaches towards alleviating them are discussed. Finally, this thesis is placed into the defined context, and its history and outline are given.

1.1 Automatic Speech Recognition

Figure 1.1 shows a block diagram of an automatic speech recognition (ASR) system. It shows three major modules: (1) feature extractor, (2) acoustic model, and (3) decoder. The feature extractor provides salient feature vectors at regular time intervals. The resulting features are aimed at characterizing the linguistic information of the speech signal and discarding all other sources of variability. They are then used in the acoustic model, which typically produces a measure of similarity between each temporal

feature vector and the relevant speech units (classes). This measure most often corresponds to the probability (or likelihood) of a temporal feature vector belonging to a certain class. Finally, the (global) decoder is responsible for the temporal alignment and integration of these (local) similarity measures, thereby also taking account of lexical, grammatical and possibly even semantic constraints defined by the language model.

However, it needs to be clarified that the representation of ASR in those few distinct modules is rather arbitrary and adapted to the point of view on ASR we adopt in the framework of this thesis. Different representations of ASR system architectures may be found e.g. in Gold and Morgan (2000), Huang, Acero, and Hon (2001) and Bilmes (1999a). For example, before feature extraction, one might want to add distinct modules for signal acquisition and analog-digital conversion. In spite of their importance for the quality of an ASR system, these issues are out of the scope of this thesis and are therefore not discussed here. Consequently, as the starting point of our ASR system, we take the digitized waveform for granted. Similarly, we are here not concerned with higher level language modeling (which includes dictionary, grammar, and possibly even semantics). However, the language model is implicitly considered during decoding and therefore displayed in the figure.

The figure might also suggest that the modules are distinct and well-defined, and that e.g. a certain feature extractor could be replaced by another while leaving the rest of the ASR system unchanged. However, this is not necessarily the case, as the subsequent modules might be adapted to the output of the feature extractor - or the other way around, e.g. when features are searched for which satisfy certain constraints (assumptions) imposed by the subsequent modules. In fact, an ideal feature extractor could render other system modules superfluous (or at least much simpler). While this is far from reality, it is true that there is some overlap between the different modules. As an example, speech recognition improvements in the case of environmental noise might be achieved at the level of the acoustic features and/or through adaptation of the acoustic model.

In the following, we will give a short introduction to what could be called a standard speech recognition system. By this we mean readily available and widely employed techniques which fit quite nicely in the three basic ASR system modules shown in Figure 1.1. There is a wide variety of feature extraction techniques, of which the ones which are most commonly used as well as those directly relevant to this thesis are discussed in Section 1.2. Most of today's ASR systems use either Gaussian Mixture Models (GMM) or Artificial Neural Networks (ANN) for the acoustic modeling, the former of which will be used throughout this work and is therefore briefly explained in Section 1.3. Finally, virtually all ASR systems are based on Hidden Markov Models, as introduced in Section 1.4, for the temporal decoding part.

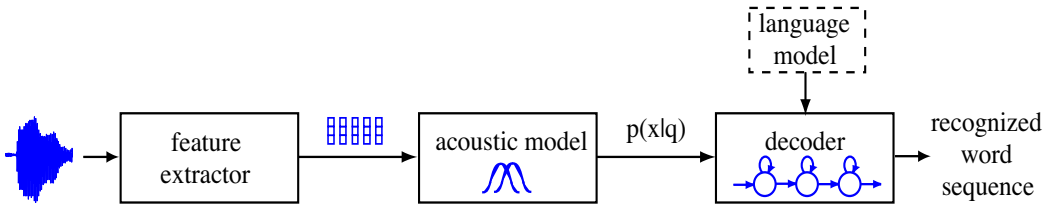


Figure 1.1: Important modules of an automatic speech recognition system.

1.2 Features for Automatic Speech Recognition

The goal of feature extraction for ASR is to provide representation of speech which permits to distinguish between the different sounds of a language, but which is at the same time insensitive to all non-linguistic variations such as speakers' characteristics and environmental influences and distortions (e.g., noise). That means that these features should be relatively stable for different examples of the same speech units, even if pronounced by different speakers and in different conditions. A cepstral representation of the speech signal, where knowledge about the human auditory system has been incorporated in the feature extraction process, seems to be dominant in state-of-the-art ASR systems (Gold and Morgan, 2000) and will be described in the following. Moreover, some extensions, variants and alternative speech representations will be discussed.

1.2.1 Mel-Frequency Cepstral Coefficients (MFCCs)

The mel cepstrum (Davis and Mermelstein, 1980) is one of the most widely employed signal representations in ASR. The exact details of how MFCCs can be calculated are covered in much of the ASR literature (Gold and Morgan, 2000; Rabiner and Juang, 1993; Huang, Acero, and Hon, 2001) and need not concern us here. For us, it is sufficient to know that spectral analysis is followed by an integration of the power spectrum within about 20-26 triangular, overlapping filters which are equally spaced along the Mel scale. The Mel scale is perceptually motivated. It is supposed to model the sensitivity of the human ear, which has been shown to distinguish far better between close sounds in low than in high frequencies (Huang, Acero, and Hon, 2001; Rabiner and Juang, 1993). The Mel scale can be approximated by the following equation (Young et al., 1995):

$$Mel(f) = 2595 log_{10}\left(1 + \frac{f}{700}\right). \tag{1.1}$$

The resulting filter magnitudes are then pre-emphasized (to approximate the unequal sensitivity of human hearing at different frequencies) and logarithmically compressed (to model the power law of human hearing). Finally, an orthonormal transformation (usually the Discrete Cosine Transform, DCT) is applied to calculate MFCCs. Typically, only the first 13 MFCCs are used for ASR, including the 0-th coefficient as a measure of the energy.

1.2.2 Temporal Derivatives

Features like MFCCs as discussed above provide a good smooth estimate of the local short-term spectrum (Gold and Morgan, 2000). However, one of the dominant characteristics of the speech signal is its dynamics. In fact, temporal changes were shown to be important for human speech recognition (Huang, Acero, and Hon, 2001). Also for ASR, they can represent discriminant information, which might not be sufficiently captured in the so-called static cepstral coefficients described above. The conventional way to include information on these dynamics is to augment the static feature vectors with temporal derivatives. First and second order temporal derivatives (also called delta and acceleration coefficients, denoted by D and A respectively) can be estimated by using regression equations (Young et al., 1995). The feature vector of 13 MFCCs is thus augmented by 13 delta and 13 acceleration coefficients, resulting in an overall feature vector dimension of 39. MFCCs are widely employed, and often serve as a reference in order to measure performance improvements of newly developed features (Bilmes, 1999a; Gales and Young, 1996; Garner and Holmes, 1998; Holmes, 2000; Hunt and Lefebvre, 1989; Kermorvant and Morris, 1999; Macho et al., 1999, McCourt, Vaseghi, and Harte, 1998; Nadeu, Hernando and Gorricho, 1995; Okawa, Bocchieri, and Potamianos, 1998; Wassner and Chollet, 1996; Welling and

Ney, 1998; just to name a few examples). The fact that there is such a multitude of publications in this area indicates that, although MFCC (most often including temporal derivatives) seem to be a kind of "reference" features, there is still room for improvements. This is especially true in unmatched conditions, like under different environmental noises and in applications where significant speaker variations can be expected. In the following, we will discuss two frequently applied approaches towards alleviating the effects of such variations and thus providing a higher robustness to adverse conditions.

1.2.3 Feature Transformations

One way to improve standard ASR features such as MFCCs is to apply an additional transformation to the original features (Hunt and Richardson, 1990; Haeb-Umbach and Ney, 1992). These transformations are based on certain statistics of the training data. Generally, they seek to orthogonalize the features and reduce the feature vector dimension, while preserving or improving class separability (and finally recognition performance). Two common examples of such transformations are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) (Fukunaga, 1972; Bishop, 1995, Duda and Hart, 1973).

PCA, also referred to as Karhunen-Loéve transformation, basically finds the directions of maximum variance of a given feature set in an unsupervised way. This transformation can be based on the calculation of the eigenvalues and eigenvectors of the global covariance matrix of all data, regardless of the class labeling. For this reason, the transformed features might well represent the signal, but not be optimal for discrimination.

In contrast, LDA aims directly at improving discrimination, relying on a measure of class separability to derive a transformation. Therefore, the class labeling for the training data needs to be known. Typically, we want to maximize the inter-class variance while minimizing the intra-class variance. To achieve this, a linear transformation can be derived according to different criteria, such as the trace of the ratio between inter- and intra-class covariance matrices (Fukunaga, 1972). One such transformation is thus given by the eigenvectors of this covariance ratio matrix. As the above criterion should be maximized, those eigenvectors whose eigenvalues are largest should be chosen for dimensionality reduction (as is also the case for the PCA described above).

1.2.4 Noise Reduction Techniques

Another way to make standard ASR features more robust is to apply noise reduction techniques such as spectral subtraction (SS) and cepstral mean normalization (CMS) (Gold and Morgan, 2000; Huang, Acero, and Hon, 2001). These methods typically try to remove an estimate of the noise from the speech signal, thus reducing a possible mismatch between training and testing conditions in the feature space. While SS is designed to cope with additive noises (e.g., car noise, office noise), CMS handles convolutional distortion (e.g. microphone or telephone distortions, reverberations).

SS is based on the assumptions that the speech signal has been corrupted by (rather stationary) additive noise, and that the clean signal and the additive noise are uncorrelated. In order to obtain the power spectrum of the clean signal, an estimate of the noise power spectrum (typically obtained during non-speech periods) is subtracted from the power spectrum of the corrupted signal.

CMS removes the (long-term) average of the cepstrum from each cepstral feature vector. This processing is based on the fact that convolutional disturbances in the time domain become additive in the cepstral domain. As this subtraction in the cepstral domain corresponds to a division in the spectral domain, this technique is also called Cepstral Mean Normalization.

1.2.5 Alternative Representations

Although MFCCs are widely employed, they are by far not the only powerful speech representation in terms of ASR performance. Another example, which is very similar to MFCCs as it is based on similar processing steps, is perceptual linear prediction (PLP), and different processing steps of the two approaches may even be combined to yield other variants of features (Gold and Morgan, 2000). Likewise, the techniques described in Sections 1.2.2 to 1.2.4 simply represent common examples amongst a multitude of possibilities, which might be combined directly or in a different, often more sophisticated form with other feature extraction approaches. As an example, RASTA-PLP (Hermansky and Morgan, 1994) is a modified version of PLP processing which is based on filtering temporal trajectories of subband energies. It is thus related to temporal derivatives (Section 1.2.2) and can also be interpreted as a short-time version of CMS (Section 1.2.4) (Gold and Morgan, 2000).

For some applications, it might be desirable to use features in the spectral domain. This is for instance the case in "missing data" processing (Cooke et al., 2001), as well as in the HMM2 approach which will be introduced in Chapter 4. However, depending on the kind of model used in subsequent ASR modules, common spectral representations of speech are often not competitive with features in the cepstral domain. An exception to this are the recently developed Frequency Filtered Filterbank coefficients (FF) (Nadeu, 1999; Nadeu, Macho and Hernando, 2001). Frequency filtering can be applied directly to Mel frequency filterbank coefficients (obtained at an intermediate stage during the extraction of MFCCs, as explained in Section 1.2.1). Frequency filtering consists of simply calculating the difference between two coefficients of the same feature vector, e.g. $x(f) - x(f-1)$ for first order FF coefficients and $x(f+1) - x(f-1)$ for second order ones (in which case they are denoted as FF2). One of the advantageous effects of this frequency filtering is a decorrelation of the coefficients, while staying in the spectral domain. In addition, FF and FF2 features yield competitive speech recognition results to MFCCs, which makes them particularly attractive features for the applications mentioned above.

An alternative spectral representation of speech is given by wavelet coefficients. A possible advantage of these features is their property of providing information on different resolution levels in time as well as in frequency. In fact, they offer a better temporal resolution at higher frequencies and a better frequency resolution at lower frequencies. Wavelet related features have been used in different ways in ASR (e.g., Kadambe and Srinivasan, 1994; Wassner and Chollet, 1996; Long and Datta, 1996; Long and Datta, 1998; Kryze et al., 1999; Farooq and Datta, 2001), but as yet have not been shown to be competitive with state-of-the-art features.

While features like MFCC are derived using knowledge about human speech perception, alternative representations consider speech production related information. Of these, formants (defined as the resonance frequencies of the vocal tract) are a compact and highly efficient representation of the time-varying characteristics of speech (Rabiner and Juang, 1993), which are supposed to be robust in noise. Formants have been shown to be useful especially for vowel classification. In ASR, they have been used in combination with other state-of-the-art features. However, one of the drawbacks of formants lies in the difficulty of reliably estimating them, and as yet they are not widely used as features in ASR systems.

1.3 Gaussian Mixture Models (GMM)

There are a number of different ways to implement the acoustic model. Of these, the most widely employed are Artificial Neural Networks (ANNs) (Bourlard and Morgan, 1994, Bourlard and Bengio, 2002) and Gaussian Mixture Models (GMMs) (Rabiner and Juang, 1993). While both these methods

exhibit certain advantages and drawbacks, they show a comparable performance (provided that suitable (possibly different) features are used for either of these two approaches). GMMs can be seen as the classic approach, moreover offering a suitable framework for the issues investigated in this thesis. Therefore, we will focus on the use of GMMs for (local) acoustic modeling.

In ASR, GMMs are typically used to model the distribution of the data belonging to a certain class (e.g., phoneme or sub-phone unit, represented by the HMM states). GMMs are universal approximators of densities, i.e., given a sufficient number of mixture components, they can approximate any distribution (Bishop, 1995; Huang, Acero, and Hon, 2001). For the case where the covariance matrices are diagonal, as considered here, the associated probability density function is defined as follows:

$$p(x) = \sum_{g=1}^{G} c_g \prod_{f=1}^{F} \frac{1}{\sqrt{2\pi\sigma_{gf}^2}} \cdot e^{-\frac{1}{2}\left(\frac{x^f-\mu_{gf}}{\sigma_{gf}}\right)^2}, \tag{1.2}$$

where G is the number of mixture components and c_g are the weights associated to the respective mixtures, F is the number of (scalar) components x^f of a feature vector x (corresponding to the feature vector dimension d), and μ_{gf} and σ_{gf}^2 are the means and variances of all mixture components $g = 1 \ldots G$ and all feature vector components $f = 1 \ldots F$. The mixture weights c_g are positive and

$$\sum_{g=1}^{G} c_g = 1 . \tag{1.3}$$

GMMs can be trained using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). As we will discuss EM training for Hidden Markov Models (where the state probability distributions are modeled by GMMs) in Section 1.4.3, EM training for GMMs will not separately be discussed here.

1.4 Hidden Markov Models

Hidden Markov models (HMMs) can be seen as a generalization of GMMs, which are suitable for sequential data. In ASR, (first order) HMMs are typically used to represent the density of sequences of T acoustic vectors $X = x_{1,T} = \{x_1, x_2, \ldots, x_t, \ldots, x_T\}$, as shown in Figure 1.2. The basic idea underlying HMMs is to introduce a hidden (unknown) variable q_t describing the state of the system at time t, and to factor the density of a sequence into several more simple terms: the initial state probabilities $P(q_0)$, the state transition probabilities $P(q_t|q_{t-1})$, and the emission probabilities $p(x_t|q_t)$. As these HMM state emission probabilities are here represented by GMMs, we refer to the entire system as Gaussian Mixture Hidden Markov Model (GM-HMM).

1.4.1 Notation

Basic notations used throughout this document are defined below and visualized in Figure 1.2:

- x_t is the observed vector at time step t,
- q_t the HMM state at time t, where Q is a path through the HMM,
- $p(x_t|q_t)$ is the HMM emission probability, where the instantiation $p(x_t|q_t = i)$ is the probability to emit x_t in state i,
- $P(q_0)$ is the initial state probability of the HMM,

- $P(q_t|q_{t-1})$ is the HMM state transition probability, where the instantiation $P(q_t = i|q_{t-1} = j)$ is the probability to go from HMM state j at time $t-1$ to state i at time t,
- N is the number of HMM states,
- T is the size of the sequence $x_{1,T} = \{x_1, x_2, \ldots, x_T\}$.

The likelihood of the data sequence $X = x_{1,T}$ given the model parameters θ at training step k is then

$$L(X|\theta) = p(x_{1,T}|\theta^k). \tag{1.4}$$

1.4.2 HMM Assumptions

In the HMMs presented in this section, we assume that the value the hidden (state) variable takes is governed by a first order Markov process. The observations (feature vectors) then depend on the resulting assignment of this variable. We can thus formulate two conditional independence assumptions, regarding transition and emission probabilities. Firstly, it is assumed that the state q_t is conditionally independent of any preceding variables given the previous state q_{t-1}:

$$P(q_t = i|q_{t-1} = j, q_{1,t-2}, x_{1,t-1}) = P(q_t = i|q_{t-1} = j). \tag{1.5}$$

This equation is in fact a generalization of the first order Markov assumption. Moreover, it is assumed that the transition probabilities are independent of time, i.e., they depend only on the origin j and the destination i. Secondly, the probability of emitting x_t at time t depends only on the state $q_t = i$ and is conditionally independent of the past states and observations:

$$p(x_t|q_t = i, q_{1,t-1}, x_{1,t-1}) = p(x_t|q_t = i). \tag{1.6}$$

This equation is frequently referred to as the output-independence assumption.

1.4.3 EM Training

Supposing that a sequence of acoustic vectors has been generated by a hidden Markov model, the underlying sequence of HMM states is generally not known. Therefore, the data observed is said to be "incomplete" and consequently, HMM parameters can not be estimated directly. The Expectation Maximization (EM) algorithm offers a way to circumvent this problem, using an iterative two-step procedure.

The goal of the EM algorithm is to maximize the likelihood $L(X|\theta)$ of the data X, given the model parameterized by θ. EM solves the problem of the incomplete data by introducing hidden variables such that the knowledge of these variables would simplify the learning problem. Hence, in the first step of each iteration (referred to as E-step), the values of these hidden variables are estimated, while in the

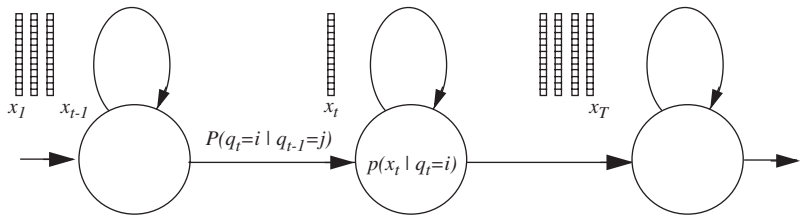


Figure 1.2: Left-right Hidden Markov Model (HMM).

second step (referred to as M-step), the expectation of the log likelihood of the observations and the hidden variables is maximized, given the previous values of the parameters. This two-step process is repeated iteratively and is proved to converge to a local optimum of the likelihood of the observation (Dempster, Laird & Rubin, 1977).

The adaptation of EM suitable for HMMs is also referred to as "Forward-Backward" or "Baum-Welch" algorithm (Baum and Petrie, 1966; Baum et al., 1971), which is briefly outlined in the following[1]. As stated before, for the case of HMMs the hidden variables correspond to the state assignments, i.e. the sequence of states Q. Introducing an indicator variable $Z = \{z_{i,t}\}$ such that $z_{i,t}$ is defined to be 1 when $q_t = i$ and 0 otherwise, the joint likelihood of the observations and the hidden variable is then given by:

$$L(X, Q) = P(q_0)\prod_{t=1}^{T}\prod_{i=1}^{N}\left[p(x_t|q_t = i)^{z_{i,t}}\prod_{j=1}^{N}P(q_t = i|q_{t-1} = j)^{z_{i,t}\cdot z_{j,t-1}}\right]. \tag{1.7}$$

We further define the following auxiliary function:

$$A(\theta|\theta^k) = E_Q[\log L(X, Q|\theta)|X, \theta^k]. \tag{1.8}$$

As stated before, the E-step consists of computing the expectations of the hidden variables, given the current parameters and the data. This can be done using a recursive estimation of some intermediate "forward" and "backward" variables (which explains the name Forward-Backward algorithm). In the M-step, given the estimated values of the hidden variables and the data, new parameters (i.e., for the transition probabilities and emission distributions) maximizing A are found. Thus, at the k-th iteration, one computes

$$\theta^{k+1} = \underset{\theta}{arg\,max}\, A(\theta|\theta^k). \tag{1.9}$$

It can be shown that maximizing A also maximizes the likelihood of the data $L(X|\theta)$ (Dempster, Laird, and Rubin, 1977).

1.4.4 Decoding

The aim of HMM decoding is to find the sequence of HMM states which best explains the input data, while at the same time taking account of phonological, lexical and syntactical constraints in the case of ASR. Therefore, under the typical HMM assumptions (see Section 1.4.2), the recognized word sequence can be obtained by finding the path Q^* which maximizes the joint likelihood of the data and the hidden variables, given the model parameters:

$$Q^* = \underset{Q}{arg\,max}\left[P(q_0)\prod_{t=1}^{T}[p(x_t|q_t)P(q_t|q_{t-1})]\right]. \tag{1.10}$$

This is usually done with the Viterbi algorithm (Viterbi, 1967), which is based on a recursion quite similar to the calculation of the forward variable mentioned above[2].

[1] A detailed description of this algorithm can be found in Section 3.2.3.

[2] This issue will be discussed in more detail in Section 3.2.4.

1.4.5 Automatic Speech Recognition using HMMs

In automatic speech recognition, HMMs are used to model the different speech units. Typically, these speech units correspond to phonemes, and each phoneme model comprises several states (often connected in a left-right topology with loops, as shown in Figure 1.2). Words can be considered as concatenations of phonemes. Hence, to obtain a word model, the relevant phoneme models can be concatenated to form one large HMM (where the dictionary defines the sequence(s) of phonemes of which a word may be composed). Similarly, sentences are sequences of words, and a sentence model corresponds to a concatenation of word models. For this case, the grammar can define possible sentences, or, alternatively, the probability of a certain sequence of words.

It is usually not possible to train each phoneme model separately, as the phoneme boundaries are not known (and accurately segmenting a database in terms of phonemes is a hard and highly time-consuming task). Therefore, "embedded training" can be used, where one large HMM is created as the concatenation of the phoneme models corresponding to the pronounced word or sentence. This typically only requires a transcription (i.e., sequence of words) and a dictionary. All phoneme models can then be trained simultaneously using the EM algorithm described in Section 1.4.3.

The aim of speech recognition is to find the sequence of pronounced words, given an acoustic sequence. This is done by searching for the sequence of words, and thus phonemes and states, which best explains the input data. For this, the Viterbi algorithm (briefly introduced in Section 1.4.4) is usually applied. Obviously, the resulting sequence of states is constrained to be of the same length as the input sequence to be recognized. Further constraints are provided through dictionary and grammar (and possibly even semantics), which can be considered parts of the language model and are highly dependent on the application. An adapted definition of the language model can significantly reduce the search space during decoding, making speech recognition more accurate and more efficient. However, these aspects of ASR will not be further investigated in this thesis.

In the following, we will again focus our attention on HMMs, and discuss particularly their weaknesses and limitations.

1.4.6 Weaknesses

HMMs are quite powerful statistical models which can in principal model any probability distribution over sequences. As seen above, efficient training and decoding algorithms are available, so that they have become a standard in automatic speech recognition. However, it is often claimed that HMMs suffer from a number of limitations. Potential (often cited) weaknesses of standard HMMs (as discussed in this chapter) with respect to their application to ASR, include

- a poor modeling of acoustic context (in fact, each observation is assumed to be conditionally independent of the past, given the current HMM state, and therefore all the context should be reflected in the assignment of the (discrete) state variable),
- the assumption that speech can be well represented by a succession of steady segments (represented by the states) with instantaneous transitions between them (in fact, the observations are assumed to be identically distributed, given the HMM state),
- a poor modeling of duration (in fact, duration is primarily modeled by the transition probabilities and therefore supposed to follow an exponential distribution, and, moreover, the contribution of these transition probabilities to the overall likelihood score is often negligible).

However, it can be argued (Bilmes, 1999a) that these weaknesses[3] are generally due to practical constraints, rather than caused by the model itself. They are not inherent properties of HMMs, but result from the way HMMs are used. Bilmes confirmed that in general, an HMM can accurately model any real-world probability distribution, given a sufficiently large number of hidden states and a sufficiently rich class of observation distributions. However, for practical problems, this typically implies a large number of model parameters, necessitating a prohibitive amount of training data.

As a consequence of the ever-persisting problem of limited training data, we have to deal with yet another weakness:

- a constrained model topology and number of parameters (typically chosen a priori).

Naturally, this problem is aggravated if there is a lot of variability in the data. For speech, this is typically the case. Sources of variability include additive noises and channel distortions, but also speech and speaker variations such as voice quality, context, stress, speaking rate and style (Junqua and Haton, 1996). Although some of these distortions and variations can effectively be removed from the data (e.g., using noise reduction techniques as mentioned in Section 1.2.4), state-of-the-art ASR systems still suffer from

- a limited robustness in adverse conditions.

The higher the variability of the data, the higher the amount of data and the complexity of the model that is needed for an adequate data representation. Given that the amount of available training data is limited, HMMs have to be constrained. For the case of ASR, this typically results in left-right phone (or triphone) models with a limited number of states, using a limited number of Gaussian mixtures to estimate the observation probabilities. These restricted class of HMMs is then likely to exhibit the first set of weaknesses listed above.

Finally, another potential weakness of standard HMMs is that

- the model is usually not trained using a discriminant criterion (i.e., minimizing the classification error (Duda and Hart, 1973)), but instead using a maximum likelihood criterion (i.e., maximizing the likelihood of the data associated with each speech unit).

Much research has been devoted to overcome at least some of these limitations (and this thesis is yet another attempt). Some of the most important achievements are described below.

1.5 Alternative Models

In the last section, we have outlined some of the weaknesses of the application of HMMs in the framework of ASR (which of course may equally apply to other real-world problems). It was stated that the main reason for these limitations lies in the constraints which have to be imposed on HMMs (due to a the problem of limited training data and a high variability in the data). While illustrating this statement with some examples, we will introduce attempts to overcome some of these limitations for the case where training data is limited.

With an unlimited amount of training data (including all potential sources of variability), large HMMs could be trained which could effectively approximate the real data distribution. A more realistic way to deal with this problem is to train not only the parameters of a given model, but also learn the

[3] As well as other supposed / frequently criticized weaknesses (Bilmes, 1999a), which are however closely related to the ones outlined here.

model structure from data. One attempt in this direction is the concept of Buried Markov Models (BMMs, Bilmes, 1999a; Bilmes, 1999b). BMMs are built upon the conventional hidden Markov model approach. In fact, extensions are added to standard HMMs where these were found to be deficient in the framework of a particular task. These extensions take the form of conditional dependencies, e.g. between components of different feature vectors, which are added based on a conditional mutual information criterion. This also has a potential to resolve other HMM weaknesses. Observations are no longer considered to be conditionally independent of the past, but the most salient dependencies, e.g. of preceding observations, are explicitly considered. Bilmes showed that these additional dependencies improved the model[4].

These BMMs can be interpreted as Dynamic Bayesian Networks (DBNs). DBNs can be seen as a generalization of conventional HMMs (Smyth, Heckerman and Jordan, 1997), providing the means for incorporating additional dependencies (such as described above), but also for dealing with (possibly hidden) auxiliary information (e.g., articulator positions) (Zweig, 1998; Stephenson, 2003). As compared to the HMM framework, which is based on an observed sequence of feature vectors resulting from a hidden sequence of HMM states, DBNs allow the integration of additional observed or hidden variables, and the consideration of statistical dependencies between all these variables.

BMMs and DBNs may be also seen as a more sophisticated and data-driven extension to "conditional dependent HMMs", "correlation HMMs" or "conditionally Gaussian HMMs", where it is assumed that the local probability depends not only on the state but also on the previous frame(s) (Huang, Acero, and Hon, 2001; Wellekens, 1987; Ostendorf, Digalakis, and Kimball, 1996). Related to this are also "Segmental HMMs" (Gales and Young, 1993), where the observations are conditionally dependent not only on the current HMM state, but also of the mean of a "segment" of speech to which they belong. These models are an example of a class of approaches which is referred to as "segment models" (Ostendorf, Digalakis, and Kimball, 1996), which deal with sequences of frames, rather than with independent frames. Besides segmental HMMs, they also include linear dynamical systems and other types of trajectory models.

Other (perhaps more traditional) ways to relax the conditional independence properties of HMMs include the use of temporal derivatives (see Section 1.2.2) (or information from longer time scales, as will be discussed later in this thesis) as additional feature vector components in order to broaden the scope of each temporal feature vector (Huang, Acero, and Hon, 2001; Bilmes, 1999a).

To improve duration modeling with standard HMMs, one can use several HMM states featuring the same local probability distribution. Depending on the connectivity of these states, a minimum duration can be modeled (as often applied in the framework of HMM/ANN), and, if these states have self-transitions, the resulting duration distribution is a sum of geometric distributions, which is a much richer model than the geometric distribution implicitly modeled by one self-looped HMM state. An alternative way of alleviating the duration problem is to use higher order HMMs, where the state transition probabilities depend on the n previous states (for n-th order HMMs). However, an n-th order HMM can be transformed into an equivalent first order HMM (Jelinek, 1997; Bilmes, 1999a; Huang, Acero, and Hon, 2001). To improve duration modeling in standard HMMs, duration can be incorporated into HMMs (Wang, 1997), e.g. via an explicit, state-dependent duration distribution (Rabiner and Juang, 1993, Huang, Acero, and Hon, 2001, Russel and Moore, 1985; Levinson, 1986). Also in segment models as

[4]In spite of the fact (which he had proven previously) that even standard HMMs can represent correlation between feature vectors and (related) information about the acoustic context is represented indirectly via the hidden (state) variable.

described above, explicit duration models are incorporated for each state (Ostendorf, Digalakis, and Kimball, 1996).

As already mentioned above, the state emission probabilities of an HMM can be modeled by an ANN. These systems are referred to as (hybrid) HMM/ANN. HMM/ANNs have several advantages over the standard GM-HMMs. They do not need strong assumptions about the distribution of the input data, and they can approximate any kind of non-linear discriminant functions (Bourlard and Morgan, 1994; Bourlard and Bengio, 2002). Another way to improve discrimination in ASR systems (including standard GM-HMMs) is to choose training methods which maximize posterior probabilities (by adjusting the parameters of all models simultaneously), instead of maximizing the data likelihood (as done in EM training), such as maximum mutual informations (MMI) (Bahl et al., 1986, Bilmes, 1999a).

Also, ANNs can be used at the level of the decoder, thus replacing HMMs. However, some mechanism needs to be introduced in order to ensure temporal alignments, and finally segmentation and classification. This can be realized with variants of recurrent neural networks (including a feedback of the hidden and/or output units to the input layer) and time delay neural networks (where output units are only activated after a complete speech segment has been processed) (Huang, Acero, and Hon, 2001).

As stated above, a major weakness inherent to virtually all current ASR systems is their poor robustness in adverse conditions. Systems often fail as soon as testing conditions differ from training conditions. More specifically, speech is often distorted by noise. Some noise reduction techniques have already been discussed in Section 1.2.4. Instead of, or in addition to, reducing the noise in the features, the effects of noise on ASR performance can be alleviated by using models which can implicitly handle these difficult conditions. Examples of this kind include the "multi-stream" approach (Morris, Hagen, Glotin, and Bourlard, 2001; Hagen, 2001) where two (or more) feature streams (ideally containing complementary information) are processed independently up to some stage, before being recombined. The recombination can take place at different stages in the recognition process, using different methods and possibly sophisticated weighting techniques which reflect some reliability measure of the different streams. Related to multi-stream are approaches like "factorial HMMs" (Ghahramani and Jordan, 1997), "HMM decomposition" (Varga and Moore, 1990; Varga and Moore, 1991) and "parallel model combination" (Gales and Young, 1995). The idea common to these methods is the modeling of several independent sources (e.g., speech and noise) by different HMMs. However, these models can be seen as special cases of large conventional HMMs (Bilmes, 1999a).

As a particular example of multi-stream, the speech signal can be decomposed into several frequency sub-bands (referred to as "multi-band" approach) (Bourlard and Dupont, 1996; Hagen, 2001). As these sub-bands are processed independently, band-limited noise present in one sub-band would not affect the other sub-bands. An extension to multi-band processing based on Markov random fields was investigated in (Gravier, Sigelle and Chollet, 2000). Also related to these methods is the "missing data" approach, where so-called missing data masks are calculated (e.g., based on the local signal-to-noise ratio (SNR) estimates), and recognition is based only on data which is supposed to be reliable (Cooke et al., 2001).

For most of the techniques described above, limited improvements were often only possible at the price of a substantial increase in computational complexity, and/or for some specific (artificial) task. However, to the best of our knowledge, these approaches have generally not shown substantial improvements over the conventional HMM approach, so that standard HMMs remain (in spite of their obvious limitations) the most often applied model in automatic speech recognition (Huang, Acero, and Hon, 2001).

1.6 Correlation in GMMs and HMMs

In the previous sections, we have introduced GM-HMMs, the assumptions imposed by them, their limitations, as well as ways to overcome these limitations. We have stated that, in spite of a variety of possible alternatives, the classical GM-HMMs remain very competitive. Given these models, in the following we will focus on the particular aspect of correlation modeling. It is clear that data from natural processes (such as speech) has some redundancy, and therefore is necessarily correlated. This thesis exploits different ways of modeling this correlation, as explained below.

Let us first consider the different types of correlation one has to deal with in ASR. When looking at a spectral representation of speech, one can distinguish between correlation in time and correlation in frequency. At the level of the final features used in ASR (such as MFCCs), this translates to correlation between different feature vectors and correlation within each feature vector[5], denoted as inter- and intra-frame correlation.

Because modeling correlated data typically requires more parameters (and thus more training data), one might prefer to reduce the correlation inherent in the speech signal. This can be done during some preprocessing steps, e.g. converting the highly correlated spectral representation of speech into the more decorrelated Mel-cepstrum, or applying LDA. Typically, these techniques significantly reduce (but do not completely eliminate) the intra-frame correlation. However, substantial inter-frame correlation persists. In the following, we discuss how persisting intra-frame correlation as well as inter-frame correlation can be dealt with in GMMs and HMMs respectively.

In the context of standard GM-HMMs, modeling of correlation is performed at two explicit levels. While correlation within each temporal feature vector is modeled by the GMMs, correlation between feature vectors is modeled by the HMM. As stated before, these models are theoretically able to represent any data (sequence) distribution- including correlated data. However, the limitations of the models become evident in real-world applications.

In GMMs, correlation can be modeled through the combination of the different mixture components. However, a large number of Gaussians is needed to model correlated (and higher-dimensional) features. In practice, this is demonstrated by severe performance losses when using (correlated) spectral data as compared to data in the cepstral domain, given the same number of Gaussian mixtures. However, as seen above, the number of Gaussians is limited due to the limited amount of training data. Therefore, it is often not possible to use a large enough number of parameters as would be required for the modeling of this correlated data.

In HMMs, modeling of correlation is implicitly done through the model topology. It was shown that, contrary to common criticism, HMMs can represent correlation between feature vectors (Bilmes, 1999a). However, as in the case of GMMs, a large number of model parameters (here: HMM states) is needed.

In the framework of this thesis, different ways of dealing with correlation are exploited. In particular, we investigate the effects of shifting the modeling of correlation further towards GMMs or HMMs respectively. On the one hand, contextual information can be included in each feature vector, and the correlation between the components of this enhanced vector is then to be modeled by the GMM. On the other hand, each vector can be split up into smaller sub-vectors. In this case, correlation within these sub-vectors is still modeled by (lower-dimensional) GMMs, but the correlation between different sub-vectors can be modeled by an HMM. These issues will be addressed in Chapters 2 and 3.

[5]Although the information in one frame does not necessarily cover only one time step.

1.7 Thesis History and Outline

The arrangement of this thesis follows to a certain extent the chronological order in which the research was carried out. Firstly, Chapter 2 deals with what could be called "early ideas". In fact, this thesis started off with a preliminary study on including information from longer time scales (additionally to temporal derivatives) into each acoustic feature vector. This method proved to be beneficial for the speech recognition performance, especially in the case of noise. This motivated us to go one step further and consider different levels of resolution not only in the temporal, but also in frequency dimension. For this, wavelets were chosen as features, as they are inherently multi-resolutional in both these dimensions. However, the dimension of wavelet feature vectors is very high, and the feature vector components tend to be highly correlated. For this reason, we searched for a model which would offer a less parameter-intensive and more powerful representation of wavelet features than the conventionally applied GMMs. Consequently, wavelet-domain hidden Markov trees (WHMT) were imported from the field of computer vision, and used for local likelihood estimation in each HMM state (thus replacing the GMMs).

However, while research on this WHMT-HMM system was put on hold after a short time and at a very preliminary stage (at which some potential of this approach became apparent, but could not yet be underpinned with positive speech recognition results), the underlying concept gave rise to further (even more promising) investigations on using a more general kind of models than WHMTs (or GMMs) for the local likelihood estimations in each HMM state. Thus, the idea of HMM2 was born, which rapidly became the focus of research carried out in the framework of this thesis. In particular, a certain instantiation of HMM2, featuring a bottom-up topology with only few (self-looped) states to model the frame likelihoods, was investigated and constitutes the kernel of this thesis. Consequently, the largest part of this document deals with this HMM2 system and issues related to it. Chapter 3 is entirely devoted to the HMM2 theory, while Chapter 4 presents practical aspects as well as ASR results.

When analyzing the mechanisms underlying HMM2, it was discovered that this system implicitly extracts pertinent information from the speech features, which could in turn be used as features for conventional HMMs. Chapter 5 investigates the use of HMM2 as feature extractor. The resulting features are called "HMM2 features". Again, positive results were obtained for the recognition of noisy speech. As it was assumed that the most promising HMM2 features bear a resemblance with the formant representation, this issue is investigated in more depth in Chapter 6, where they are compared to "true" (as well as to other automatically extracted) formant features in terms of their performance on a vowel classification task. Once again, encouraging results were obtained.

Of course, the research effectuated in the framework of this thesis does not exhaust the subject. Chapter 7 summarizes the results obtained so far and outlines promising directions for future research.

The work accomplished in the framework of this thesis and reported in this document was generally done by its author, of course under the guidance of thesis director and advisors. Exceptions to this are explicitly mentioned. The author's work also resulted in or contributed to several publications (see the list of publications, page 118). Publications which are primarily based on the work of the author of this thesis are [1], [3], [4], [7], [8], [9], [10], [14]. To publications [12] and [15], both their first and second authors made major contributions. While the contributions of the author of this thesis to the other publications are of a lesser importance, they are also cited for the sake of completeness and where their contents is related to the subjects dealt with in the respective sections of this document.

Chapter 2

Modeling Time/Frequency Correlation

This chapter focuses on some very preliminary work, which formed the starting point of this thesis. In this context, two different research directions were investigated: the multiple time scale feature combination approach, and the wavelet-domain hidden Markov tree approach. In the multiple time scale feature combination approach, features obtained from longer time scales are appended to conventional feature vectors, and the augmented feature vectors are processed in GM-HMMs in the usual way. In the wavelet-domain hidden Markov tree approach, features obtained from different (temporal and frequency) resolutions are investigated in combination with a particular HMM, which models the correlation between the different components of a feature vector. It is shown that this approach can also be integrated into the HMM paradigm. While some positive results for both these methods are reported, it should be noted that these early research directions were not explored comprehensively. However, they inspired the new ideas which will be described in the main part of this thesis.

Before giving a brief overview of these two preliminary approaches, we will first introduce the general experimental setup used throughout this thesis.

2.1 General Experimental Setup

This section gives some information about the general setup used for the experiments reported in this thesis. In particular, we will briefly discuss database specifications and evaluation criteria. While the choice of databases was primarily motivated by our research goals, an additional consideration was conformity with other ongoing research (Kermorvant, 1999; Glotin, 2000; Hagen, 2001). This also had an influence on some more decisions concerning the practical experimental setup, e.g., parameters such as frame rate and analysis window size during feature extraction, the division of data in different independent sets for training and testing, and the significance tests. These issues will be discussed in the following.

2.1.1 Databases

The majority of the experiments reported in this thesis have been done on the OGI Numbers corpus which was released in 1995 (Cole et al., 1995). This database (in the following referred to as Numbers95) is a multi-speaker telephone speech database in American English. As a collection of naturally produced connected numbers it contains e.g., zip codes, numbers from addresses, birth dates and phone numbers, with their orthographic transcriptions. The vocabulary consists of 30 words, and no grammar is defined (i.e., each word may follow any other with equal probability).

Also, some experiments on Numbers95 data degraded by additive noise will be reported. The noises were partly drawn from the Noisex database (Varga et al., 1992), and partly provided by Daimler-Chrysler in the framework of a common European project (SPHEAR). Three different noises were artificially added at four different signal-to-noise ratios (SNR). The SNR provides a measure for the amount of noise added to the speech signal. The desired ratio between the speech signal S and the noise N (over all frequencies f) was obtained for each sentence (excluding silence parts) through adaptation of the gain factor g, using following equation (Hagen, 2001):

$$SNR = 10 \cdot log_{10} \frac{\sum_f S^2(f)}{\sum_f gN^2(f)}. \tag{2.1}$$

However, it should be noted that the SNR might not be the optimal measure of degradation of the speech signal in terms of the effects of the noise on automatic speech recognition performance. I.e., for some noises, relatively good speech recognition results might be achieved at low signal-to-noise ratios, while for others, severe degradations are observed already at a relatively high SNR. For instance, the effects of white noise are typically more severe in terms of performance degradations than those of band limited or low frequency noise at the same SNR. Consequently, for a given noise and SNR, more importance should be given to the performance comparison between two methods than to results reported for one method alone.

In correspondence with (Kermorvant, 1999; Hagen, 2001), two independent subsets of the Numbers95 database were used. The training set consisted of 3590 utterances, comprising approximately three hours of (clean) speech. Generally, optimization was done on this set. The test set[1] consisted of 1206 utterances, comprising 4670 words. The same utterances, artificially corrupted with the different noises as described above, were also used for testing in noise.

If not stated otherwise, feature extraction was done with a frame rate of 12.5ms on 32ms analysis windows. Training was started on 27 3-state left-right monophone models, where the emission probabilities were modeled with single Gaussian distributions. While the number of Gaussians was increased (up to mixtures of ten Gaussians) in successive training steps, EM training was carried out. Finally, 80 triphone models were generated from the monophones, followed by a final parameter re-estimation using the EM algorithm.

Some preliminary testing was also done using the 1998 release of the Aurora database (Pearce, 1998). The aim of this Aurora project was to compare systems (in particular front-ends) of different research institutes. For this reason, the kind of models, the training algorithm, and other options were specified in this project and left unchanged for our experiments. The Aurora database was derived from the TIDigits database and includes artificially added noise. It has a small vocabulary (11 words), pronounced by different speakers. While the characteristics of this database are fairly similar to those of Numbers95, Aurora was also conceived for multi-condition training, i.e. training under different noise conditions, which was used in the experiments reported here.

A third database, which was used for some particular vowel classification (and not speech recognition) experiments, is the American English Vowels (AEV) database. Details on AEV will be given in Section 6.3.1.

[1] In fact, this test set corresponds to the test set used in (Kermorvant, 1999), and is a super-set of the one used in (Hagen, 2001).

2.1.2 Evaluation Criterion

The evaluation criterion for all the speech recognition experiments is the word error rate (WER), calculated using the following equation:

$$WER = \frac{ins + del + subs}{total} \cdot 100\% , \tag{2.2}$$

where ins, del and $subs$ are the numbers of insertions, deletions and substitutions respectively, and are found by comparing the recognizer output with the correct transcription (Young et al., 1995). $total$ is the total number of words, given the correct transcription. The significance of performance improvements are evaluated using confidence intervals (CI), which are estimated using the following equation:

$$CI = WER \pm z_{\alpha/2} \sqrt{\frac{WER \cdot (100 - WER)}{total}} , \tag{2.3}$$

where $z_{\alpha/2} = 1.96$ for $\alpha = 0.05$, corresponding to the 95% confidence interval[2] (McClave and Sincich, 2000; Bronstein, 1989).

2.2 Multiple Time Scale Feature Combination

While a lot of progress has been made during the last decades in the field of ASR, one of the main remaining problems is that of robustness. Typically, state-of-the-art ASR systems work very efficiently in well-defined environments, e.g. for clean speech. However, their performance degrades drastically under different conditions. As discussed in Chapter 1, many approaches have been developed to circumvent this problem. Here, we investigate the influence of using additional information from relatively long time scales to noise robustness.

In state-of-the-art ASR systems, feature extraction techniques analyze the speech waveform and produce an acoustic vector, a representation of the speech signal suitable for further processing by HMMs. Typically, this analysis is performed on rather short windows (up to about 30ms) of the speech signal. Some contextual information is provided by appending derivatives to the original feature vector. However, there is no information covering a longer time span, e.g., spanning the length of one syllable.

Recently, it has been shown that this kind of long-term information could improve robustness of ASR systems. For examples, TempoRAl Patterns (TRAPs) (Hermansky and Sharma, 1999) use additional information of up to one second. Information regarding syllables is considered in (Wu et al., 1998). In the same spirit, we are here using features covering relatively long time scales, and which are combined with conventional feature vectors.

2.2.1 Additional Features from Longer Time Scales

Usually, a feature vector contains only information obtained from an analysis window of about 30ms length as well as first and second order temporal derivatives. This feature vector's context is modeled entirely by the topology of the HMM. To introduce some additional contextual information into each feature vector, we appended new features, obtained either by analysis over a longer time span or by averaging over a number of subsequent feature vectors, as a second, time-synchronous, stream. For

[2]Again, this significance test corresponds to the one in (Kermorvant, 1999; Hagen, 2001). It should however be noted that it is based on assumptions (such as Gaussianity of the WER) which are often not fulfilled (Bronstein, 1989; McClave and Sincich, 2000; Mokbel, 1992).

example, we computed the average feature vector over 9 vectors, looking at four adjacent frames on either side of a feature vector, thus covering a time span of 112ms in total for our second time scale. This corresponds roughly to the amount of contextual information frequently employed in hybrid HMM/ANN systems (Bourlard and Morgan, 1993), as well as to approximately half the length of a syllable. Other experiments used even longer time spans of up to 2s for the additional information stream.

The features obtained from each time resolution can be seen as separate information streams, and the multiple time scale feature combination approach is thus a particular kind of multi-stream processing (Hagen, 2001). Different streams can be treated synchronously or asynchronously (Mirghafori, 1999). Here, we only consider the synchronous processing of the streams, where their combination is straight-forward and can be done either on the feature level or on the level of the local likelihoods (referred to as feature combination or likelihood combination respectively, as described e.g. in (Okawa, Bocchieri, and Potamianos, 1998)). In the case of feature combination, the feature vectors from different streams are combined to form a single feature vector, and the HMM state likelihoods are then calculated in the usual way (e.g., using equation 1.2). In the case of likelihood combination, equation 1.2 can be employed at the level of each stream, and the resulting stream likelihoods $p(x^s)$ are the combined as follows:

$$p(x) = \prod_s p(x^s)^{w^s}, \tag{2.4}$$

where w^s is the weight associated with stream s.

In the experiments reported in this section, we used feature combination, i.e., features calculated using windows covering different time spans of the speech signal were combined to form a single feature vector.

2.2.2 Preliminary Experiments

The effects of introducing features from a second time scale (using feature combination) were tested on the Aurora database (see Section 2.1.1). Conventional MFCCs with their first and second order derivatives were used for the first time scale. Analysis windows had a length of 32ms, and features were extracted every 10ms. Noise reduction techniques (i.e., spectral subtraction and blind equalization, Kermorvant, 1999) were applied. Our Aurora baseline system yields an average WER (for 0-20dB) of 13.4%, with a 95% confidence interval of [13.2%.13.6%]. It should be noted that, apart from changing the feature extraction part (and thereby also the feature vector and model dimension), the same parameters as defined in the Aurora specification were used throughout the tests, which might not be optimal for some cases.

Experiments were run where the features of the second time scale were obtained through averaging 9 subsequent features of the first time scale. Therefore, the resulting features span more than 100ms, and are of the same dimension as those from the first time scale (resulting in large feature vectors). Results are shown in Table 2.1. The average word error rate of 12.7% achieved by our multiple time scale system can be considered a significant improvement, given the baseline results and the 95% confidence interval mentioned above.

More tests have been carried out, e.g., calculating conventional MFCCs for the second time scale features, but on a window of 128ms (yielding an average WER of 12.5%); taking the average over 17 frames (WER= 13.1%); and taking the average over 201 frames which corresponds to roughly 2 seconds (WER=37.7%). The last experiment was repeated, but for the second time scale only one coeffi-

cient (the energy) was calculated and appended to the original feature vector. This way, a WER of 12.3% was obtained (see Table 2.2), which is the best result obtained on all our multiple time scale systems. The same setting, but only regarding a window of one second, yielded a WER of 13.7%.

	Noise1: Exhibition Hall	Noise2: Babble Noise	Noise3: Train	Noise4: Car moving	Average of Noises 1..4
Clean	1.5	1.7	1.5	1.5	1.5
20 dB	2.1	3.0	2.0	1.7	2.2
15 dB	3.7	6.4	2.8	1.6	3.6
10 dB	8.0	15.4	5.0	2.3	7.7
5 dB	17.4	34.1	11.6	4.7	17.0
0 dB	36.0	58.6	27.5	10.5	33.2
-5 dB	66.8	78.6	53.4	29.0	57.0
Average 0..20dB					12.7

Table 2.1: Word error rate on the Aurora database: Two time scales, the second time scale being the average over 9 subsequent frames of the first time scale and consisting of 13 coefficients.

	Noise1: Exhibition Hall	Noise2: Babble Noise	Noise3: Train	Noise4: Car moving	Average of Noises 1..4
Clean	1.5	1.7	1.5	1.1	1.5
20 dB	2.2	2.6	1.5	1.0	1.8
15 dB	3.7	5.5	2.5	1.3	3.3
10 dB	7.5	13.8	4.7	2.3	7.0
5 dB	14.5	31.7	11.5	4.3	15.5
0 dB	36.7	60.0	28.0	10.5	33.8
-5 dB	70.9	80.1	59.7	29.4	60.0
Average 0..20dB					12.3

Table 2.2: Word error rate on the Aurora database: Two time scales, the second time scale being the average over about 2s of the energy coefficients of the first time scale.

Figure 2.1 shows the performance of the multiple time scale systems from Tables 2.1 and 2.2 in comparison to the baseline. The relative WER ratio

$$RR = \frac{WER(baseline) - WER(2timescales)}{WER(baseline)} \cdot 100 \tag{2.5}$$

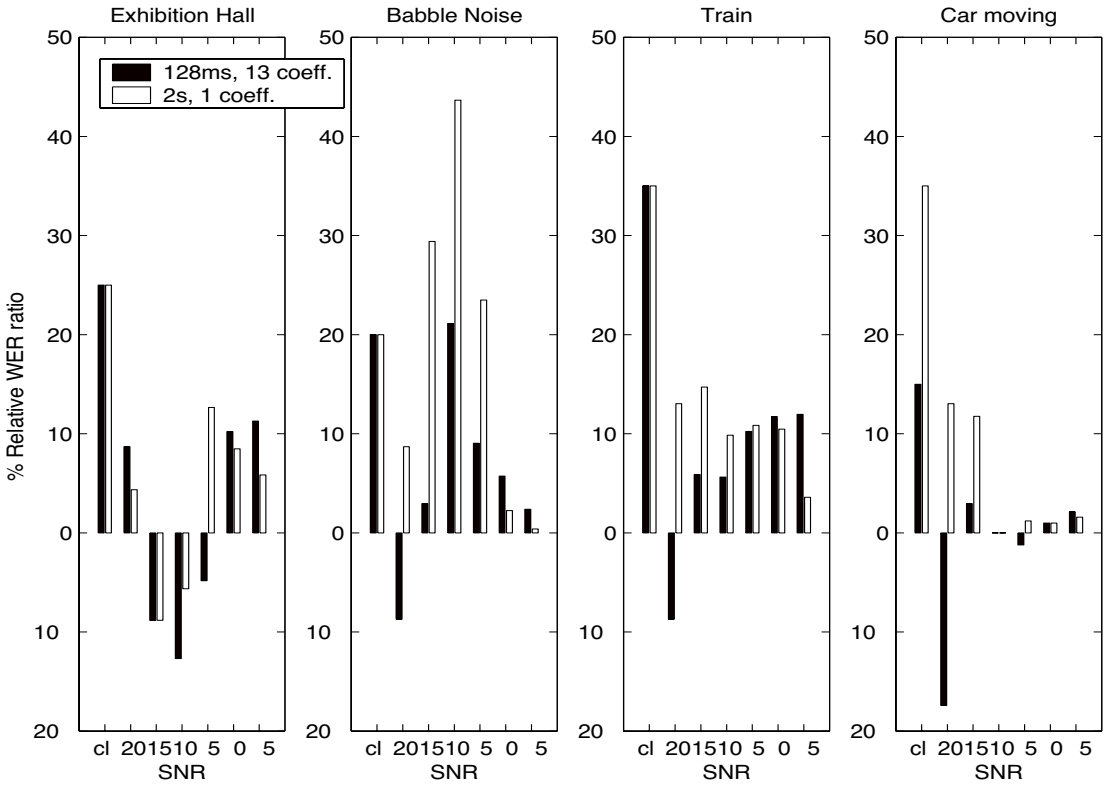


Figure 2.1: Aurora database: Percentage of relative WER ratio of the multiple time scale systems (from Tables 2.1 and 2.2) as compared to the baseline. Positive values mean a decrease in WER, i.e., a better recognition performance.

is visualized, with positive values meaning a decrease in WER (and thus better performance) for the multiple time scale system. It can be seen that in most cases both multiple time scale systems perform better than the baseline, and that the system with just one additional component for the second time scale, calculated over 2s (light bars in the figure), generally performs better.

In summary, the best tested multiple time scale system uses 14 coefficients plus their first and second order derivatives. 13 MFCCs (including energy) were calculated on 32ms of speech and one long-term energy coefficient was appended. This coefficient was obtained by averaging the energy coefficients over a time span of approximately 2 seconds, centered around the window from which the first time scale coefficients were calculated. First and second order derivatives were appended. This way, a significant improvement was gained as compared to our single time scale baseline system.

2.2.3 Discussion

In this section, the multiple time scale feature combination approach was investigated. It was shown to significantly increase robustness of ASR systems in the case of additive noise, as compared to state-of-the-art systems. However, some improvements might lead to an even better recognition performance. These include a higher number of time scales, changing lower and upper cut-off frequencies of the frequency bands used for the different time scales as well as varying the length of the analysis (or averaging) window. Also, an additional feature transformation such as LDA might prove to be advantageous. For instance, it might replace the averaging technique for the calculation of the second time scale, which

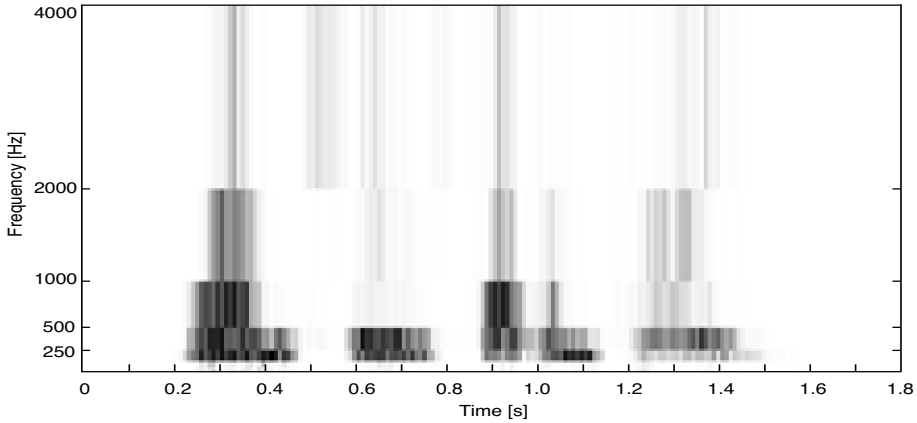


Figure 2.2: Wavelet features obtained from the Numbers95 database: The words pronounced are "one two seven three". Dark/light regions correspond to high/low energy coefficients.

was applied in the experiments described above. Different feature extraction or preprocessing techniques might be used for the various time scales. Likelihood combination may be used instead of (or in addition to) feature combination, pointing towards a more general multi-stream framework, as described e.g. in (Ellis, 2000; Hagen, 2001).

2.3 Wavelet-domain Hidden Markov Trees

Above, it has been shown that features obtained from two different time scales could improve ASR performance. In this section, we extend this idea, using features obtained on even more time scales. Multi-resolution features have been applied in ASR before, e.g. (McCourt, Vaseghi, and Harte, 1998). Here, wavelets (Daubechies, 1992) are employed, which are inherently multi-resolutional along the temporal as well as the frequency dimension. However, the focus of this section is on a new modeling technique, which is especially adapted to wavelet features. In fact, the local likelihood of a wavelet feature vector is calculated using a so-called Wavelet-domain Hidden Markov Tree (WHMT). Phoneme likelihoods obtained using the WHMT and wavelet features are then combined at the level of each temporal HMM state with those obtained by Gaussian distributions using MFCCs. This corresponds to the likelihood combination (as mentioned above) of the two feature streams.

2.3.1 Introduction

Due to their inherent multi-resolution characteristics, wavelet coefficients offer an implicit way to exploit information on multiple time scales. In fact, the time scale of the analysis varies with frequency, providing greater temporal resolution for higher frequencies, and better frequency resolution for lower frequencies. The wavelet transformation is calculated using shifted versions of a low-pass scaling function and shifted and dilated versions of a bandpass wavelet function. These functions, if chosen reasonably, form an orthonormal basis, as, e.g., the Daubechies-4 transformation (Daubechies, 1992), which was used for the experiments reported below. Daubechies-4 wavelet coefficients c_i (with $i = 0 \ldots t_w/2 - 1$, where t_w is the length of the analysis window) can be calculated using an iterative procedure. Given a signal s, the wavelet coefficients at the highest frequency resolution level can be calculated using:

$$c_i = \frac{1-\sqrt{3}}{4\sqrt{2}} s_{2i} - \frac{3-\sqrt{3}}{4\sqrt{2}} s_{2i+1} + \frac{3+\sqrt{3}}{4\sqrt{2}} s_{2i+2} - \frac{1+\sqrt{3}}{4\sqrt{2}} s_{2i+3} . \tag{2.6}$$

Similarly, the Daubechies-4 scaling coefficients a_i can be calculated using

$$a_i = \frac{1+\sqrt{3}}{4\sqrt{2}} s_{2i} + \frac{3+\sqrt{3}}{4\sqrt{2}} s_{2i+1} + \frac{3-\sqrt{3}}{4\sqrt{2}} s_{2i+2} + \frac{1-\sqrt{3}}{4\sqrt{2}} s_{2i+3} . \tag{2.7}$$

These scaling coefficients are subsequently used for the calculation of the wavelet and scaling coefficients of the next lower frequency resolution, replacing the original signal s in equations 2.6 and 2.7.

Wavelet coefficients have successfully been applied, e.g., in the field of image processing (Antonini et al., 1992; DeVore, Jawerth, and Lucier, 1992). Recent advances take advantage of inter-coefficient dependencies by modeling wavelet feature vectors with a special kind of HMM: the Wavelet-domain Hidden Markov Model (Crouse, Nowak, and Baraniuk, 1998). In fact, while the wavelet transformation is expected to decorrelate the signal, there still remain some major statistical dependencies. Particularly, adjacent coefficients in the time-frequency plane show similar behavior, as can be seen from a speech sample in Figure 2.2. It becomes obvious that coefficients are correlated across time (horizontal axis) as well as frequency (vertical axis). These correlations are inherent properties of the wavelet transform, referred to as clustering and persistency respectively.

There are three types of Wavelet-domain Hidden Markov Models, taking account of different correlations: (1) Independent Mixtures, treating each coefficient as statistically independent of all others, (2) Markov Chains, regarding only correlations across time, and (3) Markov Trees. Here, we focus on the third model, which emphasizes the dependencies within one feature vector across frequency. On the basis of (Crouse, Nowak, and Baraniuk, 1998; Choi and Baraniuk, 1999), wavelet-domain hidden Markov trees (WHMTs) are adapted for application to ASR, and integrated into a system combining the conventional HMM approach with this new technology.

2.3.2 Wavelet Features for ASR

The wavelet transformation is calculated using shifted versions of a low-pass scaling function and shifted and dilated versions of a bandpass wavelet function, which, if chosen reasonably, form an orthonormal basis (Daubechies, 1992). A way of interpreting wavelet coefficients of a speech signal is to consider their position in the time-frequency plane. Wavelet features obtained from higher frequency bands have lower frequency resolution (i.e., corresponding to larger bandpass filters) and higher temporal resolution than those obtained from lower frequency bands. Thus, the time scale of a wavelet coefficient depends on its frequency position. From this, two major properties of the wavelet transformation are apparent: locality (given the precise position of a coefficient in the time-frequency plane) and multiresolution (given the varying window size and different number of coefficients per resolution level). These characteristics could make them attractive features in the area of speech recognition. In fact, features derived from wavelet coefficients have been used in ASR, as reported e.g. in (Long and Datta, 1996; Wassner and Chollet 1996; Long and Datta, 1998; Farooq and Datta, 2001).

2.3.3 General Concepts

A wavelet feature vector can be visualized as a binary tree in the time-frequency plane, as seen in Figure 2.3. As shown in the left panel, the wavelet coefficient at the root of the tree corresponds to the lowest frequency band and the lowest temporal resolution level, spanning the whole analysis window. At the second resolution level, there are 2 wavelet coefficients, each of them spanning one half of the original

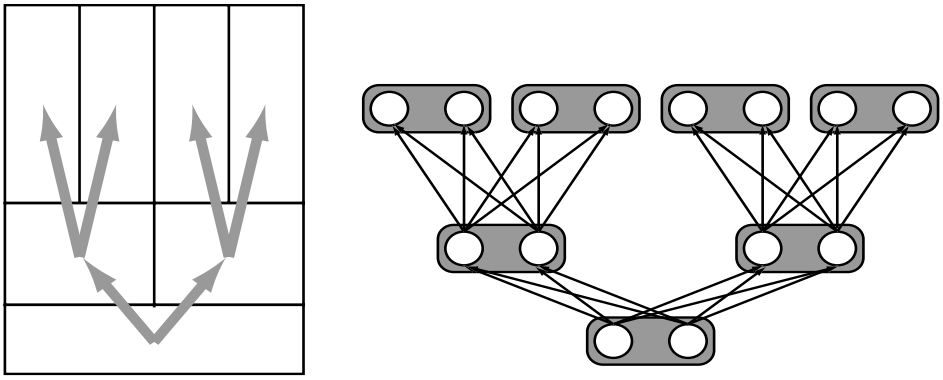


Figure 2.3: WHMT concept. The left panel shows a schema of the time-frequency plane of the lowest three resolution levels of a wavelet feature vector. The dependencies between coefficients, which are directly considered in the WHMT, are visualized by arrows. In the right panel, the corresponding WHMT is shown. Each wavelet coefficient corresponds to one node (grey in the figure), which in turn consists of two states (white circles). Transitions are introduced between the states of adjacent resolution levels, as shown in the figure.

time window. At the highest temporal resolution level R, there are 2^{R-1} coefficients, each of them corresponding to a large frequency band in a short time window.

We define a model of the structure of a binary tree (shown in the right panel of the figure), in which each "node" corresponds to one wavelet coefficient. Each node consists of two states, and a single Gaussian distribution is used to model the probability density function in each state. The choice of using two states with single Gaussian distributions was motivated by the fact that a two-state Gaussian mixture model can closely fit real wavelet coefficient data (Crouse, Nowak, and Baraniuk, 1998). We can now add connections between different nodes. Emphasizing dependencies between coefficients of adjacent frequency bands, we can construct a topology such as shown in Figure 2.3, where each state of a certain node is connected with each state of the two nodes of the adjacent resolution level (i.e., higher frequency band). Given a wavelet feature vector, the assignment of the coefficient to the different nodes in the tree is well defined. However, it is not known which state in the respective node has emitted a particular coefficient. Therefore, the state variable is hidden. In this sense, this model is a particular variant of a hidden Markov model, which we refer to as wavelet-domain hidden Markov tree (WHMT).

The parameters of this WHMT are the initial state probabilities (of the states in the root), the state transition probabilities (which are only defined for states from one resolution level to the next one along the frequency dimension, as seen in the figure), and the emission probabilities (i.e., Gaussian means and variances). Similar assumptions as made for conventional HMMs apply, and the WHMT can be trained with an adapted version of the EM algorithm (Choi and Baraniuk, 1999; Keller, Ben-Yacoub, and Mokbel, 1999).

One tree as described above might not be able to account for all potential variability in the pronunciation of a phoneme. To circumvent this problem, we can either augment the number of Gaussians in each state, the number of states in each node, or employ several WHMTs per phoneme in parallel. In our work, the third approach was used.

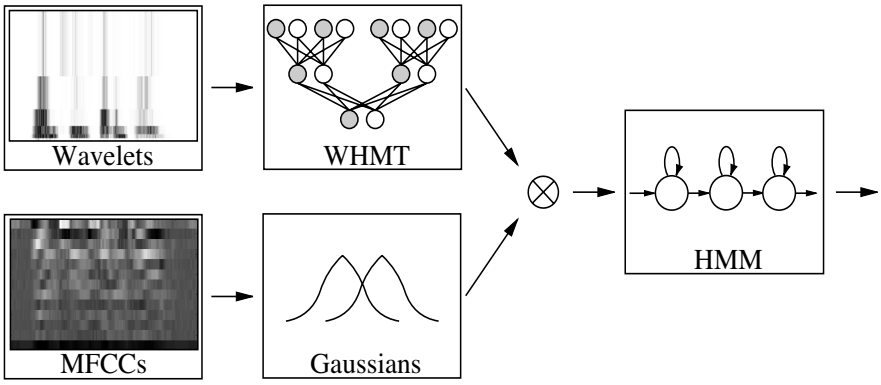


Figure 2.4: Combination of WHMTs and GMMs (taking as features wavelet coefficients and MFCCs respectively), and integration into temporal HMM system.

2.3.4 Combination with Conventional Systems

Above, we have illustrated how wavelet feature vectors could be modeled by a WHMT. In the framework of ASR, each speech unit can be modeled by one WHMT. Therefore, the WHMT model may replace the GMMs, which are conventionally used to model the distributions of the different speech units. Similar to GMMs, WHMTs can be used for the local likelihood estimation in temporal HMMs. WHMT are thus used at the level of each temporal HMM state. This HMM/WHMT system can be seen as a special mixture of HMMs.

Because of their multi-resolution properties, wavelet features have several potential advantages as compared to state-of-the-art features like MFCCs. While they have not yet been shown to outperform MFCCs for many tasks, they may provide additional, complementary information. Therefore, they may be used in a multi-stream approach as discussed in Section 2.2.1. When using WHMTs to calculate the likelihoods given a wavelet feature vector and GMMs to calculate the likelihood of the corresponding MFCC feature vector, the combination of the two streams can be done at the likelihood level (i.e., using likelihood combination, as discussed in Section 2.2.1). This is visualized in Figure 2.4.

2.3.5 Preliminary Experiments

The Numbers95 database was used for the preliminary WHMT experiments. Wavelet coefficients were calculated on 32ms windows of speech, shifted by 10ms, yielding large feature vectors of 256 components. A WHMT implementation based on (Choi and Baraniuk, 1999) was used to model these features, and one WHMT model was trained for every phoneme, (i.e., HMM state), given the hand-segmented training data. For every feature vector, the phoneme likelihoods were calculated using all these models, and a modified version of the HTK toolkit (Young et al., 1995) was used for decoding. Additionally, 13 MFCC coefficients (including energy, calculated on the same signal windows) were used to train Gaussians distributions and calculate phoneme likelihoods in a similar manner.

After preliminary testing of different kinds of wavelet transformations, the Daubechies-4 (Daubechies, 1992) transformation was chosen. Upon analyzing the phoneme confusion matrices of these tests, a high percentage of errors was observed for certain phonemes. For instance, /ey/ was very often mistaken as /ay/ or /iy/. This suggests that the employed models cannot handle the variations within these phonemes over time. By the introduction of two, three, four or six parallel WHMTs per phoneme we aimed at circumventing this problem. After an initialization based on the means of the

training data and a different (random) distortion of these means for each WHMT, training was done using an adapted version of the EM algorithm, where only the parameters of the most likely model (given the respective features) were updated. Generally, the models comprising four WHMTs performed best. Looking again at the confusion matrices, we observed that systems with different numbers of trees misrecognized different phonemes. It can be assumed that a sensible combination of these systems could be able to increase recognition performance. So we achieved some improvement in combining the two-tree and the six-tree systems by simply choosing a model on a per-phoneme bases as a function of the number of training examples. However, as defining suitable selection criteria is not a trivial task, we finally chose the four-tree system for further experiments (yielding a WER of 67.8%).

	WER
Wavelets-WHMT	67.8
MFCC-Gaussians	52.0
Combination	50.3

Table 2.3: Word error rate on Numbers95 for Wavelets-WHMTs, MFCC-Gaussians, and their combination.

In a second testing phase, we combined likelihoods calculated by our WHMT models on wavelet data with those obtained from single-Gaussian HMMs on MFCCs at the frame level. Phoneme GMM and WHMT models were trained separately. The resulting likelihoods of the two systems were combined (using likelihood combination, as described in Section 2.2.1) and then processed in an HMM system in the conventional way (Figure 2.4). This combination gave a performance improvement as compared to either of the two systems working separately, as shown in Table 2.3. In fact, the WER obtained from the wavelet data using the WHMT model was 67.8%, and the WER using the MFCCs and GMMs was 52.0%. With the combined system, a WER of 50.3% was achieved. However, although this improvement is significant (given the 95% confidence interval), it should be noted that none of these results are competitive with those obtained using state-of-the-art technology, which will be presented in the following chapters[3].

2.3.6 Discussion

We would again like to emphasize that only a very preliminary investigation of the WHMT approach was carried out. At this stage, it can not be expected to achieve competitive recognition results. However, although this matter could not be investigated in depth in this thesis, there is a lot of room for further improvement. Some possible directions for future research are outlined below.

One of the major problems of the experiments reported above is situated at the feature extraction level. In fact, data from rather low frequencies are used, which are usually disturbed by line effects for the case of telephone speech (Mokbel, Jouvet, and Monné, 1996). In particular, the four lowest levels (in frequency) of the wavelet data contain only information from frequencies below 250Hz, which probably contains no discriminant information and may even influence the recognition results in a negative way. This also causes the problem that the discriminant information from analysis windows longer than 2ms is limited. This problem is aggravated by the fact that temporal derivatives were not considered. Further-

[3]The reasons for this comparatively low performance are mainly related to a rather simple model topology (using monophone models with only one temporal state and a single Gaussian distribution, instead of triphone models with three states and mixtures of 10 Gaussian distributions, as used in later experiments).

more, there is no appropriate energy measure and no normalization. With an ameliorated signal processing, wavelet data more adapted to the characteristics of (telephone) speech could be generated.

Some problems described above are reflected at the wavelet feature modeling level. Obviously, the system relies heavily on the lower resolution levels (including the root of the WHMT model). Furthermore, a mechanism should be introduced to incorporate derivatives into our system. This could be done by adding temporal derivatives to each wavelet coefficient and modeling these small feature sub-vectors by low-dimensional Gaussians, or by introducing derivative WHMTs that are connected to the others to reflect dependencies between wavelet data and their derivatives. Some improvement might also be gained by extending the model to allow different numbers of parallel WHMTs or Gaussian mixture distributions. Moreover, the WHMT model might be made more flexible by introducing loops into the nodes/states, which, among other potential advantages, would allow phoneme duration modeling as in conventional HMMs to be directly applied to the WHMT models.

As for conventional HMMs to process MFCCs, only a rather basic system (e.g., based on single Gaussian distributions rather than GMMs) has been applied which could be replaced by a state-of-the art one. Moreover, the likelihood combination could be improved with an appropriate weighting scheme (Hagen, 2001). Furthermore, improvements could be obtained by, e.g., introducing triphones and/or more emitting states per phoneme.

More generally, the idea of modeling features by means of special HMMs working on top of the conventional HMM mechanism can be extended to features other than wavelets. For example, filterbank coefficients or even MFCCs could be modeled by double Markov chains with cross-connected hidden states. Also in this case, considering the residual correlations between adjacent coefficients of a feature vector after signal processing seems a promising research direction.

2.4 Conclusion

Above, we have presented some preliminary research on early ideas. Two different approaches were investigated: the multiple time scale feature combination approach and the wavelet-domain hidden Markov tree approach. Common to these two approaches is the use of information from different temporal resolutions (or time scales) of the speech signal in one feature vector. The correlation between the components of a feature vector was, however, modeled in different ways. For the multiple time scale feature combination approach, correlation was modeled by GMMs in the conventional way. In the WHMT approach, a new paradigm was investigated, where each feature vector was represented by a special kind of HMM, modeling correlation through the model topology.

Preliminary experiments showed some potential for both these approaches. However, there is still a lot of room for improvement. For instance, the multiple time scale feature combination approach has been further developed and has achieved even more promising results (Hagen, 2001). In the framework of this thesis, the focus of research was however directed further towards the modeling of (temporal and frequency) correlation by HMMs. Based on the experience gained from the preliminary work presented in this chapter, the HMM2 approach was developed, where a special, “secondary” HMM is used at the level of each temporal feature vector. Similar to the combination of the WHMT approach with conventional HMMs presented above, the secondary HMM works in conjunction with the usual temporal HMM. The focus of the next chapter is on the presentation of the theory on which this HMM2 approach is based.

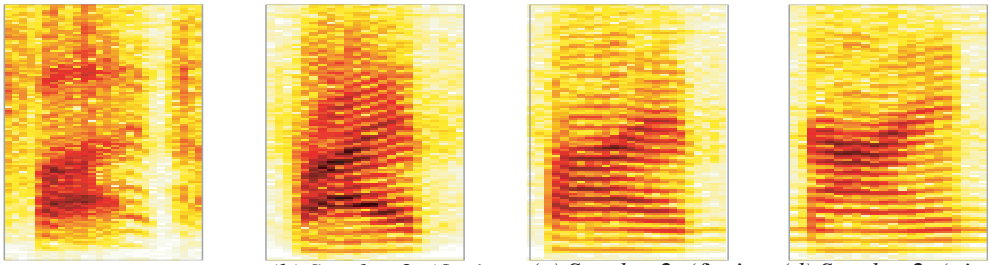
Chapter 3

HMM2: Mixtures of Hidden Markov Models

In state-of-the-art automatic speech recognition (ASR), hidden Markov models (HMMs) are widely used. While there are many suitable alternatives and design options for some parts of ASR systems such as feature extraction and phoneme probability estimation, HMMs are the uncontested model for representing temporal sequences. The success of HMMs can (at least partly) be attributed to their ability to easily accommodate temporal variations, such as different durations of phonemes, e.g. due to varying speaking rate or speakers' accents.

However, such variations do not only occur along the time axis, but can also be observed in frequency, as shown in Figure 3.1. In the spectograms depicting four different pronunciations of phoneme /ay/ (including some context), inter- as well as intra-speaker variability becomes apparent (compare Figure 3.1a with 3.1b, and Figure 3.1b with 3.1c respectively). Furthermore, Figure 3.1d shows the same phoneme pronounced in a different context, revealing the effects of coarticulation. All sub-figures suggest that the position of spectral peaks may change significantly in the time-frequency plane during the pronunciation of a phoneme.

When using HMMs, however, it is assumed that speech segments corresponding to one phoneme or sub-phone[1] unit are (1) invariant (e.g., across different speakers) enough to be modeled by the same



(a) Speaker 1: 'five' *(b) Speaker 2: 'five' (1st occurrence)* *(c) Speaker 2: 'five' (2nd occurrence)* *(d) Speaker 2: 'nine'*

Figure 3.1: Spectrograms of different pronunciations of the phoneme /ay/ by different speakers and in different contexts. Dark regions correspond to high, light regions to low energy spectral components. The vertical axis is the frequency, the horizontal one the time evolution.

[1]By this we mean the speech unit which is modeled by one HMM state.

static probability distribution and (2) stationary for their duration, which clearly is not the case. In an attempt to relax these rather rigid assumptions, and encouraged by many more practical motivations (as further elaborated in Section 3.1.2), the HMM2 approach was introduced (Weber, Bengio and Bourlard, 2000). HMM2 can be understood as an HMM mixture consisting of a primary HMM, modeling the temporal properties of the speech signal, and a secondary HMM, modeling the speech signal's frequency properties. A secondary HMM is in fact inserted at the level of each state of the primary HMM, estimating local emission probabilities of acoustic feature vectors (conventionally done by Gaussian mixture models (Rabiner and Juan, 1993) or artificial neural networks (Bourlard and Morgan, 1994)). Consequently, an acoustic feature vector is considered as a fixed length sequence of its components, which has supposedly been generated by the secondary HMM.

Although HMM2 was developed independently, a similar approach had already been proposed and used with some success in computer vision (Levin and Pieraccini, 1993; Kuo and Agazzi, 1993; Samaria, 1994; Eickeler, Müller and Rigoll, 1999). More recently, this approach was also applied to speech recognition (Werner and Rigoll, 2001). However, as further discussed below, the HMM2 approach presented here includes full EM training and was extended to take care of specificities of the problem at hand.

The purpose of this chapter is to revise theoretical and practical aspects of the HMM2 approach with regard to its application to speech recognition. Firstly, a description of HMM2 is given and motivations for applying it to speech recognition are outlined. This is followed by the HMM2 theory, including algorithms for training and decoding. Finally, a thorough analysis of HMM2, including its possible drawbacks and constraints, is given.

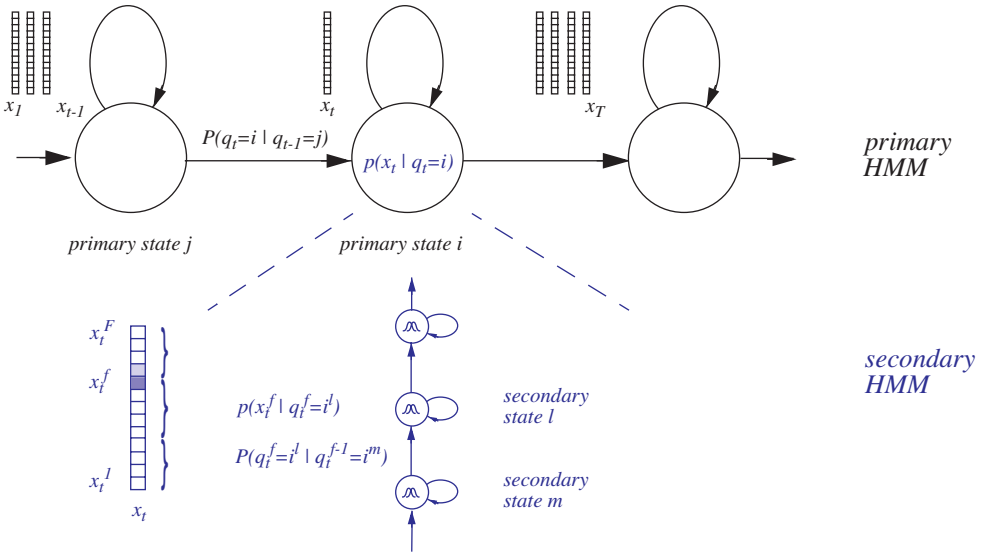


Figure 3.2: HMM2 system. In the upper part, a conventional HMM, working along the temporal axis, can be seen. The local emission probability calculation is done with a secondary HMM, working along the frequency axis (depicted in the lower part of the figure).

3.1 Introduction

3.1.1 HMM2 Description

As described in Chapter 1, HMMs are statistical models which are used to represent sequential data, e.g. a sequence of T acoustic vectors $x_{1,T} = \{x_1, x_2, \ldots, x_t, \ldots, x_T\}$ in speech recognition (as shown in the upper part of Figure 3.2). As each acoustic vector x_t can itself be considered as a fixed length sequence of its F components $x_t = x_t^{1,F} = \{x_t^1, x_t^2, \ldots, x_t^f, \ldots, x_t^F\}$, another HMM can be used to model this feature sequence (displayed in the lower part of the figure). By "component" we mean a sub-vector of low dimension. For instance, a temporal feature vector of dimension $3 \cdot F$ could be split up into F 3-dimensional sub-vectors x_t, consisting of a feature coefficient as well as its first and second order time derivatives.

While the primary HMM models temporal properties of the speech signal, the secondary, state-dependent HMM is working along the frequency dimension (supposing that a spectral data representation is used)[2]. The secondary HMM is in fact acting as a likelihood estimator for the primary HMM, a function which is usually accomplished by GMMs or ANNs in conventional systems. However, the state emission distributions of the secondary HMM are again modeled by GMMs. Consequently, HMM2 is a generalization of the standard GM-HMM system, which it includes as a particular case. In fact, a standard HMM can be realized with HMM2 in different ways, as shown in Figure 3.3. A trivial implementation of a standard HMM within the HMM2 framework is to have only one secondary HMM state which emits the entire (temporal) feature vector at once (i.e., $x_t^f = x_t$ with $F = 1$), as shown in the left panel of the figure. An alternative way of realizing a conventional HMM within the HMM2 framework is shown in the right panel. In this case, the secondary HMM consists of a number of (vertical) branches, each of which corresponds to one Gaussian mixture. The emission probability in each state is modeled by a single Gaussian distribution. Starting from the initial state, several transitions can be taken, and the

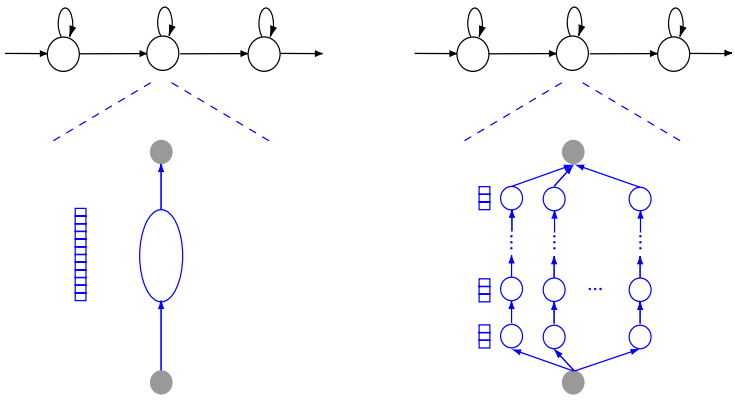


Figure 3.3: Different ways to realize a GMM within the HMM2 framework. The left panel shows the trivial solution, where the secondary HMM consists just of one state, emitting the entire temporal feature vector at once. In the right panel, each vertical branch of the secondary HMM corresponds to one Gaussian mixture component.

[2]As we focus in this thesis on the use of HMM2 for spectral data, we will use "primary HMM" interchangeably with "temporal HMM", and likewise "secondary HMM" with "frequency HMM".

associated transition probabilities correspond to the weights of the Gaussians. All following transition probabilities are set to one. Each HMM state only emits one component.

The parameters of an HMM2 are the primary HMM transition probabilities $P(q_t|q_{t-1})$, the secondary HMM transition probabilities $P(q_t^f|q_t^{f-1})$, and the secondary HMM emission probabilities $p(x_t^f|q_t^f)$.

3.1.2 Motivations

In the following, some motivations for introducing the HMM2 approach for ASR are discussed and its potential advantages as compared to conventional HMMs are outlined.

a) Better and more flexible modeling and parameter sharing

HMMs assume piecewise stationarity of the speech signal. Any signal dynamics within a segment assumed to be stationary, such as the dynamic properties of the speech signal along the feature (frequency) dimension, are disregarded. Using a secondary HMM for the local likelihood estimation, the stationarity assumption is relaxed, as a more flexible modeling of the variability and dynamics inherent in the speech signal is allowed. For instance, a spectral peak could be modeled by a single state of the secondary HMM, even though its position on the frequency axis is quite variable (as seen in Figure 3.1). Furthermore, a secondary HMM topology can be quite sparse, at the same time allowing for efficient parameter sharing. The number of parameters can be easily controlled by the model topology and the probability density function associated with the secondary HMM states.

b) Modeling of correlation through secondary HMM topology

Under the typical HMM assumptions, correlation between feature vector components is not ignored, but supposed to be modeled through the topology of the secondary HMM. Thereby, correlation of close feature vector components is emphasized in comparison to distant correlation, which corresponds to the properties of data we aim to model. In fact, HMM2 could allow a sophisticated modeling of the underlying time-frequency structures of the speech signal and model complex constraints in both the temporal and the frequency dimensions. In the same spirit, it was proposed in (Bilmes, 1999a; Bilmes, 1999b) to model time/frequency correlation in the framework of buried Markov models by using Bayesian networks to compute emission probabilities, where the connectivity of the Bayesian network was determined by the degree of mutual information between coefficients.

c) Non-linear, state dependent spectral warping

The secondary HMM automatically performs a non-linear, state dependent spectral warping. While the primary HMM does time warping and time integration, the secondary HMM performs warping and integration along the frequency axis. This frequency warping has the effect of automatic non-linear vocal tract normalization (Ikbal, Weber and Bourlard, 2002), providing a kind of unsupervised and implicit speaker adaptation (therefore tackling the problem of inter-speaker variations). Applying HMM2 in this field is also encouraged by the work of Lee and Rose (1998), who used a related frequency warping approach to speaker normalization. With the same mechanism, intra-speaker variations as well as coarticulation effects are also taken care of.

Furthermore, it could be expected that HMM2 performs a kind of implicit dynamic formant trajectory tracking. As a spectral peak (formant) can be modeled by an HMM state and a spectral valley by

another, the segmentation performed by the secondary HMM may be a good indicator for the position of a formant. Formants are assumed to carry discriminant information in the speech signal, moreover being especially robust in the case of degraded speech (Garner and Holmes, 1998, Welling and Ney, 1998).

d) Extension of multi-band processing

Currently, considerable research effort in speech recognition is being devoted to multi-band speech recognition (Morris, Hagen, Glotin and Bourlard, 2001). In this case, the full frequency band is split into multiple subbands which are processed independently (to a certain extent) by different classifiers before recombining the resulting probabilities to yield the fullband phonetic probabilities. More recently, this multi-band ASR approach was extended by using the so called "full combination approach" in which subband probability combination is performed by integrating over all possible reliable subband combinations. HMM2 can be seen as a further, more flexible extension to this approach. Indeed, all possible paths through the secondary HMM will correspond to different subband segmentations and recombinations. The frequency position of the subbands is then automatically adapted to the data, following for example formant-related structures.

The following section gives a more detailed description of the HMM2 approach, including HMM2 training and decoding algorithms.

3.2 HMM2 Theory

As stated previously, although HMM2 was proposed independently and with an entirely different motivation, it is related to similar approaches used previously for computer vision, such as Planar HMMs (Levin and Pieraccini, 1993) and Pseudo 2D HMMs (Kuo and Agazzi, 1993; Samaria, 1994; Eickeler, Müller and Rigoll, 1999). However, while these models are trained using either a planar segmentation algorithm based on Viterbi (Levin and Pieraccini, 1993), a segmental k-means algorithm (Kuo and Agazzi, 1993), or (after the two-dimensional model has been converted to a similar one-dimensional HMM) with conventional EM training (Samaria, 1994; Eickeler, Müller and Rigoll, 1999), we here develop an EM algorithm which is especially adapted to HMM2[3].

3.2.1 Notation

Basic notations used throughout this section are explained in Figure 3.2. Their definitions and some more explanations about additional notations are given below.

- x_t is the observed vector at time step t, and x_t^f is its observed component at frequency step f,
- q_t is the primary HMM state at time t, where Q is a path through the primary HMM, and q_t^f is the secondary state associated with primary state q_t at frequency step f, where Q_t is a path through the secondary HMMs associated with primary state q_t,

[3]The basics of HMM theory, including training with EM, have already been briefly outlined in Section 1.4. For the sake of completeness, and in order to make this section self-contained, certain equations will be repeated in the following.

- $p(x_t|q_t)$ is the emission probability in the primary HMM, where the instantiation $p(x_t|q_t = i)$ is the probability to emit x_t in state i, and $p(x_t^f|q_t^f)$ is the emission probability in the secondary HMM, where the instantiation $p(x_t^f|q_t^f = i^l)$ is the probability to emit component x_t^f in secondary state l of primary state i,
- $P(q_0)$ is the initial state probability of the primary HMM, and $P(q_t^0)$ is the initial state probability of the secondary HMM in primary HMM state q_t,
- $P(q_t|q_{t-1})$ is the state transition probability in the primary HMM, where the instantiation $P(q_t = i|q_{t-1} = j)$ is the probability to go from primary state j at time $t-1$ to state i at time t, and $P(q_t^f|q_t^{f-1})$ is the state transition probability in the secondary HMM associated with primary state q_t, where the instantiation $P(q_t^f = i^l|q_t^{f-1} = i^m)$ is the probability to go from secondary state m at the frequency component $f-1$ to secondary state l at frequency component f while in primary state i at time t,
- N is the number of states in the primary HMM, and N_i is the number of states of the secondary HMM associated with primary HMM state i,
- T is the size of the sequence $x_{1,T} = \{x_1, x_2, \ldots, x_T\}$, and F is the size of the sequence of components $x_t^{1,F} = \{x_t^1, x_t^2, \ldots, x_t^F\}$.

The likelihood of the data sequence $X = x_{1,T}$ given the model parameters θ at training step k is then

$$L(X|\theta) = p(x_{1,T}|\theta^k). \tag{3.1}$$

3.2.2 HMM2 Assumptions

For standard HMMs, we assume that the state sequence has been generated by a first order Markov process. For the case of HMM2, this is the case for both the temporal sequence of feature vectors, and the sequence of sub-vectors. The resulting conditional independence assumptions for transition and emission probabilities are given below.

Firstly, for the primary HMM it is assumed that the state q_t is conditionally independent of any preceding variables given the previous state q_{t-1}:

$$P(q_t = i|q_{t-1} = j, q_{1,t-2}, x_{1,t-1}) = P(q_t = i|q_{t-1} = j). \tag{3.2}$$

Similarly, for the secondary HMM it is assumed that the state q_t^f is conditionally independent of any preceding variables given the previous state q_t^{f-1}, associated with the primary state q_t:

$$P(q_t^f = i^l|q_t^{f-1} = i^m, q_t^{1,f-2}, x_t^{1,f-1}, x_{1,t-1}) = P(q_t^f = i^l|q_t^{f-1} = i^m). \tag{3.3}$$

Moreover, it is assumed that the primary and secondary transition probabilities are independent of time and frequency respectively. That means that the primary transition probabilities only depend on the origin j and the destination i. Similarly, the secondary transition probabilities only depend on the origin m and the destination l, given the primary HMM state. Secondly, for the primary HMM, the probability of emitting x_t at time t depends only on the state $q_t = i$ and is conditionally independent of the past states and observations:

$$p(x_t|q_t = i, q_{1,t-1}, x_{1,t-1}) = p(x_t|q_t = i). \tag{3.4}$$

Given the primary HMM state, the same assumption applies for the secondary HMM:

$$p(x_t^f \big| q_t^f = i^l, q_t^{1,f-1}, x_t^{1,f-1}, x_{1,t-1}, q_{1,t-1}) = p(x_t^f \big| q_t^f = i^l) \,. \tag{3.5}$$

3.2.3 Training

Since an HMM is a special kind of mixture of distributions, an HMM2, being a mixture of HMMs, can therefore also be considered as a more general mixture of distributions. As the emission and transition probabilities of the secondary HMMs are represented by mixtures of Gaussians and multinomials respectively, it should be natural that an Expectation-Maximization (EM) algorithm could be derived in a similar way as shown for GM-HMMs in Section 1.4.3. In this section, such a derivation is given for the case of HMM2, which is based on (Bengio, Bourlard and Weber, 2000).

As already discussed in Section 1.4.3, the general idea of EM is to select a set of hidden variables such that the knowledge of these variables would simplify the learning problem. Then, an iterative procedure finds a local optimum of the likelihood of the observation (Dempster, Laird and Rubin, 1977), where each iteration consists of two steps: estimation (E-step) and maximization (M-step). As shown in the following for the case of HMM2, during the E-step, the values of the hidden variables are estimated, and during the M-step, new model parameters are found, maximizing the expectation of the log likelihood of the observations and the hidden variables, given the previous values of the parameters.

In the case of HMM2, two sets of indicator variables $Z = \{z_{i,t}\}$ and $Z' = \{z_{i,t}^{l,f}\}$ are defined such that $z_{i,t}$ is 1 when $q_t = i$ and 0 otherwise, and $z_{i,t}^{l,f}$ is defined only when $q_t = i$, and is 1 when $q_t^f = i^l$, and 0 otherwise. Similar to the EM for GM-HMMs, the joint likelihood of the observations and the hidden variables is then defined as:

$$L(X, Q) = P(q_0) \prod_{t=1}^{T} \prod_{i=1}^{N} \left[p(x_t | q_t = i)^{z_{i,t}} \prod_{j=1}^{N} P(q_t = i | q_{t-1} = j)^{z_{i,t} \cdot z_{j,t-1}} \right] \tag{3.6}$$

but the emission probabilities are expressed as:

$$p(x_t | q_t = i) = P(q_t^0 \big| q_t = i) \prod_{f=1}^{F} \prod_{l=1}^{N_i} \left[p(x_t^f \big| q_t^f = i^l)^{z_{i,t}^{l,f}} \prod_{m=1}^{N_i} P(q_t^f = i^l \big| q_t^{f-1} = i^m)^{z_{i,t}^{l,f} \cdot z_{i,t}^{m,f-1}} \right] \tag{3.7}$$

Including equation (3.7) into equation (3.6) and taking the log we obtain:

$$\begin{aligned} log\, L(X, Q) = log\, P(q_0) + \sum_{t=1}^{T} \sum_{i=1}^{N} z_{i,t} \left(\begin{array}{l} log\, P(q_t^0 \big| q_t = i) + \sum_{f=1}^{F} \sum_{l=1}^{N_i} z_{i,t}^{l,f} log\, p(x_t^f \big| q_t^f = i^l) \\ + \sum_{f=1}^{F} \sum_{l=1}^{N_i} \sum_{m=1}^{N_i} z_{i,t}^{l,f} \cdot z_{i,t}^{m,f-1} log\, P(q_t^f = i^l \big| q_t^{f-1} = i^m) \end{array} \right) \\ + \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} z_{i,t} \cdot z_{j,t-1} log\, P(q_t = i | q_{t-1} = j) \end{aligned} \tag{3.8}$$

We define an auxiliary function as follows:

$$A(\theta | \theta^k) = E_Q[log\, L(X, Q | \theta) \big| X, \theta^k] \,. \tag{3.9}$$

Including equation (3.8) into equation (3.9) and moving the expectation inside the log gives:

$$A(\theta|\theta^k) = logP(q_0) + \sum_{t=1}^{T}\sum_{i=1}^{N}\hat{\gamma}_{i,t}\left(\begin{array}{l} logP(q_t^0|q_t = i) + \sum_{f=1}^{F}\sum_{l=1}^{N_i}\hat{\gamma}_{i,t}^{l,f} log\, p(x_t^f|q_t^f = i^l) \\ + \sum_{f=1}^{F}\sum_{l=1}^{N_i}\sum_{m=1}^{N_i}\hat{\xi}_{i,t}^{l,m,f} logP(q_t^f = i^l|q_t^{f-1} = i^m) \end{array}\right) \tag{3.10}$$

$$+ \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{\xi}_{i,j,t} logP(q_t = i|q_{t-1} = j)$$

with

$$\hat{\gamma}_{i,t} = E[z_{i,t}|x_{1,T};\theta^k] \tag{3.11}$$

$$\hat{\gamma}_{i,t}^{l,f} = E[z_{i,t}^{l,f}|x_{1,T};\theta^k] \tag{3.12}$$

$$\hat{\xi}_{i,j,t} = E[z_{i,t}, z_{j,t-1}|x_{1,T};\theta^k] \tag{3.13}$$

$$\hat{\xi}_{i,t}^{l,m,f} = E[z_{i,t}^{l,f}, z_{i,t}^{m,f-1}|x_{1,T};\theta^k] \tag{3.14}$$

given the model parameters θ^k at the k-th EM iteration. The expectations defined in equations (3.11) to (3.14) are calculated during the E-step of EM. Then, during the M-step, the parameters maximizing equation (3.9) are found. Thus, at the k-th iteration, we calculate

$$\theta^{k+1} = \underset{\theta}{argmax} A(\theta|\theta^k)\,. \tag{3.15}$$

As mentioned before, it can be shown that maximizing A also maximizes the likelihood of the data $L(X|\theta)$ (Dempster, Laird and Rubin, 1977).

In the following, some more details of the E-step and the M-step are given.

3.2.3.1 E-Step

Forward and Backward Variables

Let us first introduce some intermediate variables, which can be used for calculating the likelihood of a data sequence given the model, and which will also be needed for training. We define:

$$
\begin{aligned}
\alpha_{i,t} &= p(x_{1,t}, q_t = i) \\
&= p(x_t | x_{1,t-1}, q_t = i) p(x_{1,t-1}, q_t = i) \\
&= p(x_t | q_t = i) p(x_{1,t-1}, q_t = i) \\
&= p(x_t | q_t = i) \sum_j p(x_{1,t-1}, q_t = i, q_{t-1} = j) \\
&= p(x_t | q_t = i) \sum_j p(q_t = i | q_{t-1} = j, x_{1,t-1}) p(x_{1,t-1}, q_{t-1} = j) \\
&= p(x_t | q_t = i) \sum_j p(q_t = i | q_{t-1} = j) p(x_{1,t-1}, q_{t-1} = j) \\
&= p(x_t | q_t = i) \sum_j [P(q_t = i | q_{t-1} = j) \alpha_{j,t-1}]
\end{aligned}
\tag{3.16}
$$

and

$$
\begin{aligned}
\alpha_{i,t}^{l,f} &= p(x_t^{1,f}, q_t^f = i^l) \\
&= p(x_t^f | q_t^f = i^l) \sum_m [P(q_t^f = i^l | q_t^{f-1} = i^m) \alpha_{i,t}^{m,f-1}]
\end{aligned}
\tag{3.17}
$$

Thus, $\alpha_{i,t}$ corresponds to the probability of generating the sequence $x_{1,t}$ and being in primary state i at time t, and is referred to as "forward" variable of the primary HMM. Similarly, $\alpha_{i,t}^{l,f}$ corresponds to the probability of generating the sequence $x_t^{1,f}$ and being in primary state i at time t and in secondary state l at frequency f, and is referred to as "forward" variable of the secondary HMM.

Likewise, we define a "backward" variable $\beta_{i,t}$ for the primary HMM, which corresponds to the probability of emitting the sequence $x_{t+1,T}$, given that the primary state i was visited at time t:

$$
\begin{aligned}
\beta_{i,t} &= p(x_{t+1,T} | q_t = i) \\
&= \sum_j p(x_{t+1,T}, q_{t+1} = j | q_t = i) \\
&= \sum_j p[(x_{t+1} | x_{t+2,T}, q_{t+1} = j, q_t = i) p(x_{t+2,T}, q_{t+1} = j | q_t = i)] \\
&= \sum_j [p(x_{t+1} | q_{t+1} = j) p(x_{t+2,T}, q_{t+1} = j | q_t = i)] \\
&= \sum_j [p(x_{t+1} | q_{t+1} = j) p(x_{t+2,T} | q_{t+1} = j, q_t = i) p(q_{t+1} = j | q_t = i)] \\
&= \sum_j [p(x_{t+1} | q_{t+1} = j) p(x_{t+2,T} | q_{t+1} = j) p(q_{t+1} = j | q_t = i)] \\
&= \sum_j [p(x_{t+1} | q_{t+1} = j) P(q_{t+1} = j | q_t = i) \beta_{j,t+1}]
\end{aligned}
\tag{3.18}
$$

and a "backward" variable $\beta_{i,t}^{l,f}$ for the secondary HMM, which corresponds to the probability of emitting the sequence $x_t^{f+1,F}$, being in external state i at time t, and given that secondary state l was visited at frequency f.

$$
\begin{aligned}
\beta_{i,t}^{l,f} &= p(x_t^{f+1,F} | q_t^f = i^l) \\
&= \sum_m [p(x_t^{f+1} | q_t^{f+1} = m) P(q_t^{f+1} = i^m | q_t^f = i^l) \beta_{i,t}^{m,f+1}]
\end{aligned}
\tag{3.19}
$$

Consequently, the product $\alpha_{i,t} \cdot \beta_{i,t}$ corresponds to the probability of having emitted the complete data sequence $x_{1,T}$, while visiting state i at time t.

Likelihoods

The likelihood L of the sequence $x_{1,T}$ can then be calculated as follows:

$$L = \sum_i p(x_{1,T}|q_t = i) = \sum_i \alpha_{i,t} \cdot \beta_{i,t} \qquad \forall t \tag{3.20}$$

or, using only the forward variable:

$$L = \sum_i p(x_{1,T}|q_T = i) = \sum_i \alpha_{i,T} \tag{3.21}$$

Similarly, given that primary state i is visited at time t, the likelihood of x_t, corresponding to the sequence $x_t^{1,F}$, can be calculated, e.g. using the forward variable of the secondary HMM:

$$L_{i,t} = p(x_t|q_t = i) = \sum_l p(x_t^{1,F}|q_t^F = i^l) = \sum_l \alpha_{i,t}^{l,F} \tag{3.22}$$

The likelihood calculated in equation (3.22) can be used in equations (3.16) and (3.18).

Expectations

The expectation defined in equations (3.11) to (3.14) can be calculated using the intermediate variables defined in equations (3.16) to (3.19) and the likelihoods from equations (3.21) and (3.22) as follows:

$$\begin{aligned}
\hat{\gamma}_{i,t} &= E[z_{i,t}|x_{1,T}] \qquad (3.23)\\
&= P(q_t = i|x_{1,T})\\
&= \frac{p(q_t = i, x_{1,T})}{P(x_{1,T})}\\
&= \frac{p(x_{1,t}, q_t = i)p(x_{t+1,T}|q_t = i)}{L}\\
&= \frac{\alpha_{i,t} \cdot \beta_{i,t}}{L}
\end{aligned}$$

$$\begin{aligned}
\hat{\gamma}_{i,t}^{l,f} &= E[z_{i,t}^{l,f}|x_t^{1,F}] \qquad (3.24)\\
&= P(q_t^f = i^l|x_t^{1,F})\\
&= \frac{\alpha_{i,t}^{l,f} \cdot \beta_{i,t}^{l,f}}{L_{i,t}}
\end{aligned}$$

$$\begin{aligned}
\hat{\xi}_{i,j,t} &= E[z_{i,t}, z_{j,t-1} | x_{1,T}] \qquad (3.25)\\
&= P(q_t = i, q_{t-1} = j | x_{1,T})\\
&= \frac{p(q_t = i, q_{t-1} = j, x_{1,T})}{p(x_{1,T})}\\
&= \frac{p(x_{t+1,T} | q_t = i, q_{t-1} = j, x_{1,t}) p(q_t = i, q_{t-1} = j, x_{1,t})}{L}\\
&= \frac{p(x_{t+1,T} | q_t = i) p(x_t | q_t = i, q_{t-1} = j, x_{1,t-1}) p(q_t = i, q_{t-1} = j, x_{1,t-1})}{L}\\
&= \frac{\beta_{i,t} \cdot p(x_t | q_t = i) p(q_t = i | q_{t-1} = j, x_{1,t-1}) p(q_{t-1} = j, x_{1,t-1})}{L}\\
&= \frac{\alpha_{j,t-1} P(q_t = i | q_{t-1} = j) p(x_t | q_t = i) \beta_{i,t}}{L}
\end{aligned}$$

$$\begin{aligned}
\hat{\xi}_{i,t}^{l,m,f} &= E[z_{i,t}^{l,f}, z_{i,t}^{m,f-1} \,|\, x_t^{1,F}] \qquad (3.26)\\
&= P(q_t^f = i^l, q_t^{f-1} = i^m \,|\, x_t^{1,F})\\
&= \frac{\alpha_{i,t}^{m,f-1} P(q_t^f = i^l \,|\, q_t^f = i^m) p(x_t^f \,|\, q_t^f = i^l) \beta_{i,t}^{l,f}}{L_{i,t}}
\end{aligned}$$

3.2.3.2 M-Step

During the M-step, we seek to find the parameters which maximize the auxiliary function defined in equation (3.9). These parameters are

- the primary transition probabilities: a_{ij} is the probability to go from primary state i to primary state j,
- the secondary HMM transition probabilities: a_i^{lm} is the probability to go from secondary state l of primary state i to primary state m of primary state i,
- and the parameters of the probability density functions (pdf's) associated with the secondary HMM states. As noted before, we here consider these pdf's to be mixtures of Gaussian distributions with diagonal covariance matrices. However, to simplify the notation, we will give the update equations for the case of single Gaussians with diagonal covariance matrices.

We are thus looking for new parameters θ such that

$$\frac{\partial A(\theta | \theta^k)}{\partial \theta} = 0 \qquad (3.27)$$

Including equation (3.10) into equation (3.27) gives:

$$0 = \frac{\partial}{\partial\theta}\left(log P(q_0) + \sum_{t=1}^{T}\sum_{i=1}^{N}\hat{\gamma}_{i,t}\left(\begin{array}{l} log P(q_t^0 \,|\, q_t = i) + \sum_{f=1}^{F}\sum_{l=1}^{N_i}\hat{\gamma}_{i,t}^{l,f} log\, p(x_t^f \,|\, q_t^f = i^l) \\ + \sum_{f=1}^{F}\sum_{l=1}^{N_i}\sum_{m=1}^{N_i}\hat{\xi}_{i,t}^{l,m,f} log P(q_t^f = i^l \,|\, q_t^{f-1} = i^m) \end{array} \right) \right.$$
$$\left. + \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{\xi}_{i,j,t} log P(q_t = i \,|\, q_{t-1} = j) \right) \tag{3.28}$$

M-Step for primary transition probabilities

Let us first consider the transition probabilities of the primary HMM. As they are represented by multinomials, all a_{ij} are constraint to be non-negative and

$$\sum_{j=1}^{N} a_{ij} = 1 \,. \tag{3.29}$$

This constraint can be forced by introducing a Lagrange multiplier λ_i, and, instead of maximizing $A(\theta|\theta^k)$, we maximize

$$A'(\theta|\theta^k) = A(\theta|\theta^k) + \sum_{i=1}^{N}\left(1 - \sum_{j=1}^{N} a_{ij}\right)\lambda_i \,. \tag{3.30}$$

Consequently, we need to solve

$$\frac{\partial A'(\theta|\theta^k)}{\partial a_{ij}} = \frac{\partial}{\partial a_{ij}}\left(\sum_{t=1}^{T}\hat{\xi}_{i,j,t} log P(q_t = i \,|\, q_{t-1} = j) + \sum_{i=1}^{N}\left(1 - \sum_{j=1}^{N} a_{ij}\right)\lambda_i \right) \tag{3.31}$$

which gives

$$0 = \sum_{t=1}^{T}\frac{\hat{\xi}_{i,j,t}}{a_{ij}} - \lambda_i \tag{3.32}$$

Solving equation (3.32) and choosing λ_i such as to normalize the distribution, we obtain:

$$a_{ij} = \frac{\sum_{t=1}^{T}\hat{\xi}_{i,j,t}}{\lambda_i} = \frac{\sum_{t=1}^{T}\hat{\xi}_{i,j,t}}{\sum_{j=1}^{N}\sum_{t=1}^{T}\hat{\xi}_{i,j,t}} = \frac{\sum_{t=1}^{T}\hat{\xi}_{i,j,t}}{\sum_{t=1}^{T}\hat{\gamma}_{i,t}} \tag{3.33}$$

M-Step for secondary transition probabilities

The derivation of the secondary HMM transition probabilities, which are also represented by multinomials, is very similar. The respective steps are outlined below.

All a_i^{lm} are constraint to be non-negative and

$$\sum_{m=1}^{N_i} a_i^{lm} = 1 . \tag{3.34}$$

To force this constraint, a Lagrange multiplier λ_i^l is introduced, and

$$A'(\theta|\theta^k) = A(\theta|\theta^k) + \sum_{i=1}^{N}\sum_{l=1}^{N_i}\left(1 - \sum_{m=1}^{N_i} a_i^{lm}\right)\lambda_i^l . \tag{3.35}$$

is maximized. Consequently, we need to solve

$$\frac{\partial A'(\theta|\theta^k)}{\partial a_i^{lm}} = \frac{\partial}{\partial a_i^{lm}}\left(\sum_{t=1}^{T}\hat{\gamma}_{i,t}\sum_{f=1}^{F}\hat{\xi}_{i,t}^{l,m,f} log P(q_t^f = i^l \,|\, q_t^{f-1} = i^m) + \sum_{i=1}^{N}\sum_{l=1}^{N_i}\left(1 - \sum_{m=1}^{N_i} a_i^{lm}\right)\lambda_i^l\right) \tag{3.36}$$

which results in

$$0 = \sum_{t=1}^{T}\hat{\gamma}_{i,t}\sum_{f=1}^{F}\frac{\hat{\xi}_{i,t}^{l,m,f}}{a_i^{lm}} - \lambda_i^l . \tag{3.37}$$

We finally obtain:

$$\begin{aligned} a_i^{lm} &= \frac{\sum_{t=1}^{T}\hat{\gamma}_{i,t}\sum_{f=1}^{F}\hat{\xi}_{i,t}^{l,m,f}}{\lambda_i^l} \\ &= \frac{\sum_{t=1}^{T}\hat{\gamma}_{i,t}\sum_{f=1}^{F}\hat{\xi}_{i,t}^{l,m,f}}{\sum_{m=1}^{N_i}\sum_{t=1}^{T}\hat{\gamma}_{i,t}\sum_{f=1}^{F}\hat{\xi}_{i,t}^{l,m,f}} \\ &= \frac{\sum_{t=1}^{T}\hat{\gamma}_{i,t}\sum_{f=1}^{F}\hat{\xi}_{i,t}^{l,m,f}}{\sum_{t=1}^{T}\hat{\gamma}_{i,t}\sum_{f=1}^{F}\hat{\gamma}_{i,t}^{l,f}} \end{aligned} \tag{3.38}$$

M-Step for emission distributions

Let us now look at the update equations for the probability density function associated with the secondary HMM states. For the sake of simplicity, let us consider the case where the frequency sub-vectors are scalars. Then, if the emission probability of primary state i and secondary state l is defined as a Gaussian with mean μ_{il} and standard deviation σ_{il}, the log likelihood of a component is

$$\begin{aligned} log\, p(x_t^f \,|\, q_t^f = i^l) &= log\frac{1}{\sqrt{2\pi}\sigma_{il}} e^{-\frac{1}{2}\left(\frac{x_t^f - \mu_{il}}{\sigma_{il}}\right)^2} \\ &= -\frac{1}{2}\frac{(x_t^f - \mu_{il})^2}{\sigma_{il}^2} - log\,\sigma_{il} - \frac{log\,2\pi}{2} \end{aligned} \tag{3.39}$$

The update equations for μ_{il} are derived as follows:

$$\frac{\partial A'(\theta|\theta^k)}{\partial \mu_{il}} = \frac{\partial}{\partial \mu_{il}} \sum_{t=1}^{T} \hat{\gamma}_{i,t} \sum_{f=1}^{F} \hat{\gamma}_{i,t}^{l,f} \log p(x_t^f | q_t^f = i^l) \tag{3.40}$$
$$= \frac{\partial}{\partial \mu_{il}} \sum_{t=1}^{T} \hat{\gamma}_{i,t} \sum_{f=1}^{F} \hat{\gamma}_{i,t}^{l,f} \left(-\frac{1}{2} \frac{(x_t^f - \mu_{il})^2}{\sigma_{il}^2} - \log \sigma_{il} - \frac{\log 2\pi}{2} \right)$$

$$0 = \sum_{t=1}^{T} \hat{\gamma}_{i,t} \sum_{f=1}^{F} \hat{\gamma}_{i,t}^{l,f} \frac{1}{\sigma_{il}^2} (x_t^f - \mu_{il}) \tag{3.41}$$

$$\mu_{il} = \frac{\sum_{t=1}^{T} \hat{\gamma}_{i,t} \sum_{f=1}^{F} \hat{\gamma}_{i,t}^{l,f} x_t^f}{\sum_{t=1}^{T} \hat{\gamma}_{i,t} \sum_{f=1}^{F} \hat{\gamma}_{i,t}^{l,f}} \tag{3.42}$$

Similarly, the update equations for σ_{il} are derived as follows:

$$\frac{\partial A'(\theta|\theta^k)}{\partial \sigma_{il}} = \frac{\partial}{\partial \sigma_{il}} \sum_{t=1}^{T} \hat{\gamma}_{i,t} \sum_{f=1}^{F} \hat{\gamma}_{i,t}^{l,f} \log p(x_t^f | q_t^f = i^l) \tag{3.43}$$
$$= \frac{\partial}{\partial \sigma_{il}} \sum_{t=1}^{T} \hat{\gamma}_{i,t} \sum_{f=1}^{F} \hat{\gamma}_{i,t}^{l,f} \left(-\frac{1}{2} \frac{(x_t^f - \mu_{il})^2}{\sigma_{il}^2} - \log \sigma_{il} - \frac{\log 2\pi}{2} \right)$$

$$0 = \sum_{t=1}^{T} \hat{\gamma}_{i,t} \sum_{f=1}^{F} \hat{\gamma}_{i,t}^{l,f} \left(\frac{(x_t^f - \mu_{il})^2}{\sigma_{il}^3} - \frac{1}{\sigma_{il}} \right) \tag{3.44}$$

$$\sigma_{il} = \sqrt{\frac{\sum_{t=1}^{T} \hat{\gamma}_{i,t} \sum_{f=1}^{F} \hat{\gamma}_{i,t}^{l,f} (x_t^f - \mu_{il})^2}{\sum_{t=1}^{T} \hat{\gamma}_{i,t} \sum_{f=1}^{F} \hat{\gamma}_{i,t}^{l,f}}} \tag{3.45}$$

In this section, we have developed the EM algorithm for HMM2, and presented update equations for the transition probabilities of the primary and secondary HMMs and for the parameters of the emission distribution for the case where they are modeled by single Gaussian densities. In the following, we will briefly discuss decoding in an HMM2 system.

3.2.4 Decoding

The aim of HMM decoding is to find the sequence of words which best explains the input data, while at the same time taking account of phonological, lexical and syntactical constraints in the case of ASR. Therefore, under the HMM2 assumptions as discussed in Section 3.2.2, the recognized word sequence can be defined by the path Q^* which maximizes the joint likelihood of the data and the hidden variables, given the model parameters:

$$Q^* = \arg\max_{Q} \left[P(q_0) \prod_{t=1}^{T} [p(x_t|q_t) P(q_t|q_{t-1})] \right]. \tag{3.46}$$

Similarly to the calculation of the likelihood of the data given the model (as discussed above), a recursion can be used to find this path. We thus define an intermediate variable as follows:

$$
\begin{aligned}
V_{i,t} &\stackrel{\text{def}}{=} \max_{q_{1,t-1}} p(x_{1,t}, q_{1,t-1}, q_t = i) \\
&= \max_{q_{1,t-1}} \left(p(x_t | x_{1,t-1}, q_{1,t-1}, q_t = i) p(x_{1,t-1}, q_{1,t-1}, q_t = i)\right) \\
&= p(x_t | q_t = i) \max_{q_{1,t-1}} p(x_{1,t-1}, q_{1,t-1}, q_t = i) \\
&= p(x_t | q_t = i) \max_{q_{1,t-1}} \max_j p(x_{1,t-1}, q_{1,t-2}, q_t = i, q_{t-1} = j) \\
&= p(x_t | q_t = i) \max_{q_{1,t-1}} \max_j \left(p(q_t = i | q_{t-1} = j, x_{1,t-1}, q_{1,t-2}) p(x_{1,t-1}, q_{1,t-2}, q_{t-1} = j)\right) \\
&= p(x_t | q_t = i) \cdot \max_j p(q_t = i | q_{t-1} = j) \cdot V_{j,t-1}
\end{aligned}
\tag{3.47}
$$

Hence, $V_{i,t}$ is the probability of the best partial path (i.e., the most likely sequence of states) through the model for the data $x_{1,t}$, ending in state i at time t. For $t = 1 \ldots T$, $V_{i,t}$ can be computed for each state i. At the same time, for each $V_{i,t}$, the state j through which passed the best path at time $t-1$ is kept. Finally, starting from the last state (given by $i = arg\max_j V_{j,T}$), the best sequence of states can be tracked back.

At the level of the secondary HMM, the likelihood of an acoustic feature vector (i.e., a sequence of its components) given the primary HMM state $p(x_t | q_t)$ can be calculated as follows:

$$
p(x_t | q_t) = \sum_{Q_t} \left[P(q_t^0) \prod_{f=1}^{F} \left[p(x_t^f | q_t^f) P(q_t^f | q_t^{f-1}) \right] \right] \tag{3.48}
$$

This likelihood can be estimated by means of the forward recursion defined in (3.17), using equation (3.22). Alternatively, the following approximation can be used:

$$
p(x_t | q_t) \cong \max_{Q_t} \left[P(q_t^0) \prod_{f=1}^{F} \left[p(x_t^f | q_t^f) P(q_t^f | q_t^{f-1}) \right] \right]. \tag{3.49}
$$

In this case, a Viterbi recursion similar to equation (3.47) can be used at the level of the frequency HMM:

$$
\begin{aligned}
V_{i,t}^{l,f} &= \max_{q_t^{l,f}} p(x_t^{l,f}, q_t^{f-1}, q_t^f = i^l) \\
&= p(x_t^f | q_t^f = i^l) \cdot \max_m p(q_t^f = i^l | q_t^{f-1} = i^m) \cdot V_{i,t}^{m,f-1}
\end{aligned}
\tag{3.50}
$$

and $p(x_t | q_t)$ can be approximated by

$$
p(x_t | q_t) \cong \max_l V_{i,t}^{l,F} . \tag{3.51}
$$

Naturally, every term of equations 3.48 and 3.49 is conditioned on the state of the primary HMM.

Having discussed the HMM2 theory and introduced some fundamental equations, let us now investigate the HMM2 approach in yet more detail and have a closer look on the implications of the underlying mechanisms of HMM2 to data representation and discrimination.

3.3 HMM2 Data Representation

As stated before, hidden Markov models are a generalization of Gaussian mixture models (suitable for sequential data). Given a sufficiently large number of appropriately chosen parameters, these mixture models can approximate any continuous density to arbitrary accuracy (Bishop, 1995; Bilmes, 1999a). Practically, however, there are limitations. The number of parameters in a mixture model has to be appropriately chosen, depending on the task and the available training data. Furthermore, as discussed in Section 3.2.2, there are additional assumptions for the case of sequential data modeled by an HMM. Moreover, there may be constraints imposed by the HMM topology.

Naturally, in the case of HMM2, these assumptions and constraints do not only apply to the primary, but also to the secondary HMM. In this section, we investigate implications of these assumptions on the capacity of the HMM2 model for data representation, as compared to a conventional GM-HMM system.

Let us therefore consider different ways to calculate the primary HMM state likelihood: on the one hand using the conventional GMMs, on the other hand by three specific secondary HMM topologies, as described below and depicted in Figure 3.4.

- **Topology 1** (left model in Figure 3.4): Simulation of a GM-HMM with a single Gaussian distribution. The secondary model has a strict bottom-up topology without loops. The number of states is equal to the length of the sequence to be emitted. As there is only one possible state sequence, this model is a "degenerated" HMM. The local likelihoods of the secondary model states are estimated with single Gaussian distributions.
- **Topology 2** (middle model in Figure 3.4): Introduction of Gaussian mixtures (instead of single Gaussians) for the local likelihood estimation. Here, the same model topology as in Topology 1 is used, but at the level of the secondary model states, the single Gaussians are replaced by Gaussian mixtures.
- **Topology 3** (right model in Figure 3.4): Secondary HMM with loops. Compared to Topology 2, the number of states in the secondary HMM is reduced and self-transitions (loops) are added at each state. There are fewer states than emitted components, and this secondary model is a "real" HMM.

Based on these three topologies, the constraints imposed by the independent modeling of feature vector components (permitted through the output-independence assumption) and the parameter sharing (as a

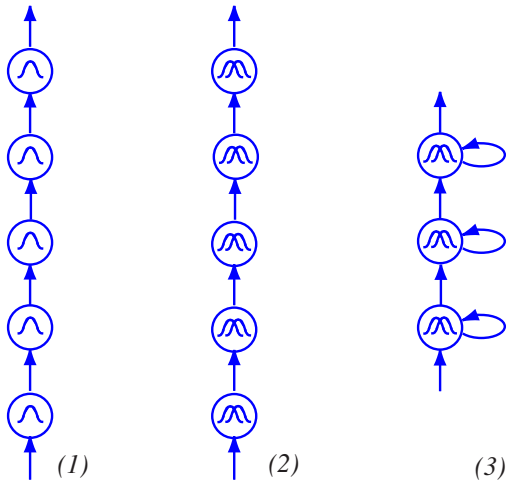


Figure 3.4: Different secondary HMM topologies.

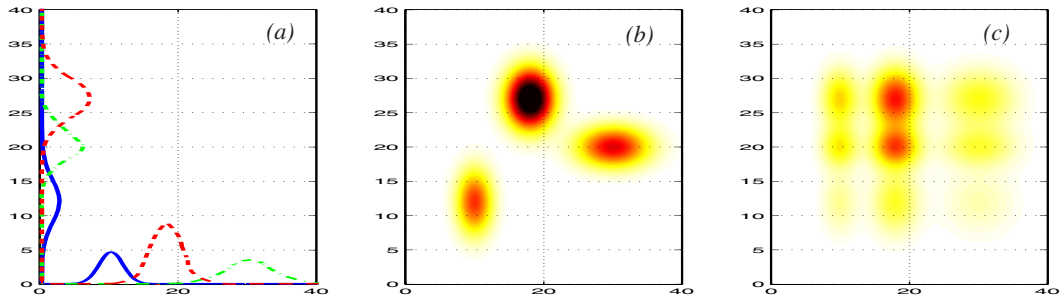


Figure 3.5: Toy example: modeling power of GMM vs. HMM. In (a), a mixture of 3 2-dimensional Gaussians is defined (i.e., Gaussian means, variances and mixture weights). This GMM is visualized in (b). In (c), a distribution resulting from an HMM (also employing the parameters defined in (a)) is shown.

consequence of the assumption of piecewise stationarity, enforced through the first-order Markov topology with fewer states than components) are discussed. It is demonstrated that the particular instantiation of HMM2 discussed here can only model a limited class of distributions, restricting the model's data representation capabilities.

3.3.1 Effects of the Independent Modeling of Components

Firstly, we will investigate the effects of independent modeling of components in the secondary HMM states, as compared to the modeling of the entire vector in a GMM. For the case of the secondary model topologies 1 and 2, equation (3.48) simplifies drastically: as there is only one possible state sequence Q_t through the model, we here deal with a "not-hidden" model. Therefore, $P(q_t^f = i^l | q_t^{f-1} = i^m) = 1$ for all transitions (i^m, i^l) defined through the model topology. For topology 1, there is even only a single Gaussian distribution, and so we obtain:

$$p(x_t|q_t = i) = \prod_{f=1}^{F} \frac{1}{\sqrt{2\Pi\sigma_{il}^2}} e^{-\frac{1}{2}\left(\frac{x_t^f - \mu_{il}}{\sigma_{il}}\right)^2} \quad \text{with } q_t^f = i^l . \tag{3.52}$$

The above equation is equivalent to the state likelihood estimation in conventional HMM systems where the distribution is modeled by a single Gaussian having F dimensions.

For topology 2 and Gaussian mixture distributions in the secondary HMM states, the simplified state likelihood equation is:

$$p(x_t|q_t = i) = \prod_{f=1}^{F} \sum_{g=1}^{G} c_{ilg} \frac{1}{\sqrt{2\Pi\sigma_{ilg}^2}} e^{-\frac{1}{2}\left(\frac{x_t^f - \mu_{ilg}}{\sigma_{ilg}}\right)^2} \quad \text{with } q_t^f = i^l . \tag{3.53}$$

This equation bears a significant difference as compared to the distribution obtained for a conventional GMM, as can be expressed using the following equation:

$$p(x_t|q_t = i) = \sum_{g=1}^{G} c_{ig} \prod_{f=1}^{F} \frac{1}{\sqrt{2\Pi\sigma_{ifg}^2}} e^{-\frac{1}{2}\left(\frac{x_t^f - \mu_{ifg}}{\sigma_{ifg}}\right)^2} \tag{3.54}$$

It can be seen that a sum of products (in the case of a GMM, equation (3.54)) has been replaced by a product of sums (in the case of a secondary HMM, equation (3.53)). Figure 3.5 shows the implications of these two equations on an example of "toy" data. It can be seen that the distribution obtained by the GMM (Figure 3.5b) is quite irregular. In fact, the shape of the distribution obtained by a GMM is practically only limited by the number of mixtures used. For example, the resulting PDF can take an (almost) elliptical form, whose principal axes are not necessarily parallel to the coordinate system. On the other hand, when modeling each feature component independently in a secondary HMM state, each mixture component in each state influences linearly all mixture components in all other states. Hence, the form of any resulting distribution is very restricted, as its principal axes inevitably follow the coordinate system's orientation. This is illustrated in Figure 3.5c. Therefore, correlation can not be modeled in the same way as in GMMs[4].

3.3.2 Effects of the Parameter Sharing

Does this drawback generalize when moving from the kind of models investigated above to hidden Markov models, or can it be compensated through some correlation modeling due to a suitable HMM topology? In the case of real HMMs (see right model in Figure 3.4), each possible path through the model corresponds to one Gaussian distribution, hence the sum over all possible paths corresponds to a Gaussian mixture (with as many mixture components as there are paths in the model):

$$\begin{aligned} p(x_t|q_t) &= \sum_{Q_t}\left[P(q_t^0)\prod_{f=1}^{F}[p(x_t^f|q_t^f)P(q_t^f|q_t^{f-1})]\right] \\ &= \sum_{Q_t}\left[\left(P(q_t^0)\prod_{f=1}^{F}P(q_t^f|q_t^{f-1})\right)\cdot\prod_{f=1}^{F}p(x_t^f|q_t^f)\right] \end{aligned} \tag{3.55}$$

where the respective products of initial and transition probabilities $P(q_t^0)\prod_{f=1}^{F}P(q_t^f|q_t^{f-1})$ represent the mixture weights.

However, if one state emits several components ($q_t^f = q_t^{f+1} = \ldots = i^l$), the underlying PDF for their data likelihood estimation is constant (i.e., the Gaussian parameters are shared for the likelihood calculation of all those components). Hence, the distributions which can be modeled by such a secondary HMM are again very restricted. This fact is depicted graphically on another "toy" example in Figure 3.6. It can be seen that the resulting distribution obeys the same restrictions as the one shown in Figure 3.5: it is not possible to model distributions whose principal axes do not follow the coordinate system's orientation. For the kind of secondary HMM we are investigating here (i.e. bottom-up topology with fewer states than emitted components), this conclusion generalizes to higher-dimensional data and a higher number of Gaussian mixtures.

In conclusion, Figures 3.5 and 3.6 both show that feature correlation can be modeled quite well by Gaussian mixture distributions, because they allow any orientation of the principal axes of the data distributions in a given coordinate system. This is not possible in the same way with a bottom-up secondary HMM with few states, because (1) the independent modeling of components in individual HMM states and (2) the parameter sharing (allowed by the stationarity assumption and enforced through

[4]It is interesting to note that the traditional multiband approach suffers from a similar handicap, for which the full-combination approach (Morris, Hagen, Glotin and Bourlard, 2001) can offer a remedy.

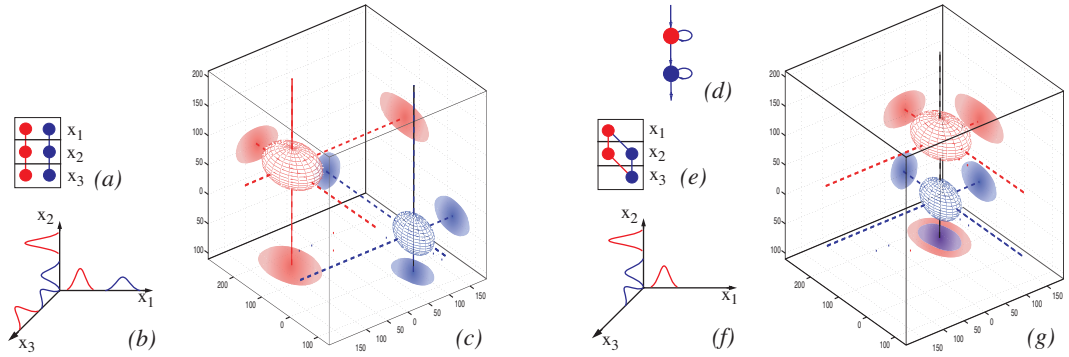


Figure 3.6: Toy example: demonstration of the modeling capacity of a GMM (left part of the figure) and a secondary HMM (right part) for the case of 3-dimensional data. The GMM consists of a mixture of 2 Gaussians with diagonal covariance matrices. The secondary HMM has 2 states as shown in (d), thus there are 2 possible paths through the model (see (e), which compares to (a) for the GMM case). In (f), the Gaussian components contributing to the resulting distribution are depicted (compare to (b) for GMM). It can be seen that, for the case of the secondary HMM, only one dimension is expanded, resulting in the distribution depicted in (g). The principal axes of this distribution are constrained to follow the axes of the coordinate system, which is not the case for the distribution resulting from the GMM (depicted in (c)).

looped HMM states) both constrain the resulting distribution to follow the orientation of the coordinate system. However, if the data conformed with the assumptions imposed through the model, HMM2 could still be appropriate. After having discussed some issues concerning data discrimination with HMM2, we will adopt a more data-driven point of view towards HMM2 and investigate the peculiarities of the speech data with respect to the above assumptions in Chapter 4.

3.4 HMM2 Data Discrimination

As stated before, the topology of the secondary HMM was chosen to be strictly bottom-up and to have fewer states than there are components in one temporal feature vector (as also seen in Figure 3.2). Therefore, each secondary HMM2 state is expected to emit a number of adjacent components, i.e. all components belonging to a certain frequency band. The number of secondary HMM states determines the number of frequency bands into which the spectrum is decomposed. The cut-off frequencies and bandwidths of these frequency bands will be dynamically determined, given the data and the model parameters, during training and decoding. These segmentations along the frequency axis could correspond to formant-like structures.

It is widely acknowledged that spectral peaks (formants) contain important discriminant information (Garner and Holmes, 1998; Welling and Ney, 1998). Therefore, the secondary HMM's frequency segmentation might represent rather discriminative information. However, HMM2 seems to suffer from the same problem as encountered in conventional HMMs: an imbalance between the contributions of HMM state likelihoods and transition probabilities to the estimation of the overall likelihood[5] (even though this effect is somewhat diminished due to the lower feature dimension in the secondary HMM). Consequently, the primary HMM state likelihoods do only insignificantly (if at all) reflect the frequency segmentation produced by the secondary HMM. The improved flexibility of the model due to the high number of paths through the frequency HMM leads to a loss of discriminability (because of the loss of information concerning formant positions), which may rule out the potential gain of frequency warping.

Under the assumption that frequency positions of different spectral regions (especially formants) represent important discriminant acoustic cues, it should be ensured that HMM2 takes them into account in an appropriate way. This problem can be solved with an additional coefficient of the feature vector, which indicates the frequency position of its respective component, as shown in Figure 3.7a (Weber, Bengio and Bourlard, 2001c). This has the effect of forcing the Viterbi algorithm to take the frequency position of each feature vector into account during the frequency segmentation.

As a toy example, Figure 3.7 illustrates the typical spectral shape of two vowel classes α and β, both consisting of 2 alternating spectral peaks (H) and valleys (L), resulting in the overall structure HLHL. These classes can be distinguished only by the position of the spectral peaks and valleys, and it is known that these positions are indeed the most important perceptual cues. Using HMM2 without frequency coefficients, the only way of modeling the differences between α and β is by the transition probabilities, which, as stated previously, do not have much influence. The two classes are therefore easily

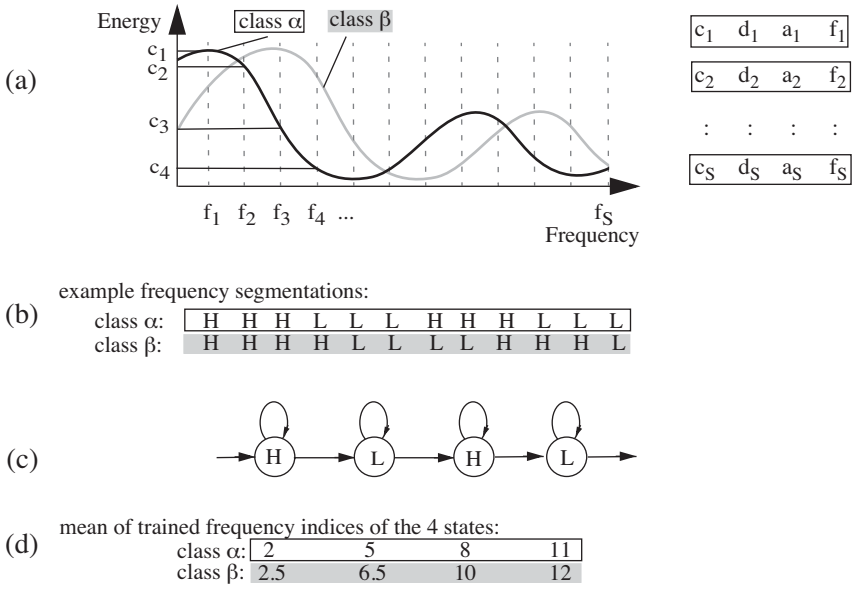


Figure 3.7: The frequency index: In (a), data assumed to be typical of the classes α and β are visualized by a black and a gray curve respectively. On the right, feature vectors (corresponding to the class α curve) as used in the secondary HMM composed of coefficients c_s, their delta d_s and acceleration coefficients a_s, as well as the frequency coefficient f_s, are shown. In (b), an example frequency segmentation is shown for each class. (c) shows a structure of an HMM with alternating H and L states, which is able to model both classes. With an additional trained frequency coefficient (as shown in (d)), discriminability can be ensured.

[5]Together with the effects of the HMM's inherent exponential duration probability distribution, this leads in conventional HMMs (as well as in our primary HMM) to a poor duration modeling. However, these problems play in the conventional case only a subordinate role. On the one hand, the poor duration modeling can be compensated for, e.g. through lexical and grammatical constraints in combination with word entrance penalties. On the other hand, the duration of a phoneme might not be an essential cue for discrimination, as this parameter varies considerably (depending on non-discriminant features such as the speaking rate).

confusable. When introducing the frequency coefficients, the Viterbi segmentation of a feature vector is in some way constrained and discriminability will be maintained. In fact, the frequency coefficient is handled in the same way as the other coefficients in a feature vector, i.e. it is modeled by the GMM. The Gaussian mean will correspond to the mean frequency of the modeled frequency band, and the variance should be an indicator of the bandwidth.

While the idea of using an additional frequency coefficient may seem surprising, it is justified in the frequency warping performed by HMM2. Improved recognition results confirm the suitability of this idea (Weber, Bengio and Bourlard, 2001c). Naturally, in standard HMMs this frequency coefficient does not give any additional information, as the frequency position of each coefficient is implicitly known from the structure.

3.5 Conclusion

This chapter has presented the motivations and foundations underlying the use of HMM2, a particular form of HMM in which emission probabilities are estimated through secondary, state-dependent, HMMs working along the acoustic feature dimension. It was shown that the parameters of this new model can be trained using the Expectation-Maximization (EM) algorithm. Including the standard multi-Gaussian HMMs as a particular case, HMM2 provides additional modeling capabilities, allowing a principled approach towards flexible modeling of the time/frequency structure of speech through warping along the temporal and frequency dimensions. However, it was shown that there are also limitations concerning the data representation and discrimination capabilities of HMM2. In the following chapter, we will investigate how these limitations might affect the practical application of HMM2 to speech recognition.

Chapter 4

Application of HMM2 as Decoder

In this chapter, the application of HMM2 as a decoder for speech recognition is investigated. After discussing the choice of features to employ in an HMM2 system, different ways of practically implementing such a model are considered. The focus of this chapter is however on the experimental evaluation. The capability of different HMM2 topologies for data representation and discrimination is evaluated for application to speech data, and compared to that of the conventionally employed GMMs. This is followed by the presentation of speech recognition experiments both for clean speech and for speech degraded by additive noise.

4.1 Experimental Setup

4.1.1 Features for HMM2

Experiments were carried out on the Numbers95 corpus (see Section 2.1.1). A major concern when working with HMM2 is the choice of the features. We investigated different representations such as filterbanks, Rasta, and MFCCs. Obviously, for the motivations outlined in Section 3.1.2 to hold, features in the spectral domain should be employed (although HMM2 might also show some advantages with different features). For most of our experiments, frequency filtered filterbanks (FF2, as explained in Section 1.2.5) were used. Compared to MFCCs, these features show only slightly worse speech recognition results on our HTK-based system (this result applies to clean data; however, performance degrades significantly in noisy conditions). In addition to staying in the spectral domain, FF2 features offer the advantage of being normalized to some degree (possibly large signal level variations are in fact smoothed out through the differencing). Twelve normalized FF2 coefficients (including one energy coefficient) were used. First and second order time derivatives were added to each feature vector.

4.1.2 HMM2 Implementation

There are different ways to implement an HMM2 systems. A straightforward realization is based on the implementation of a generalized form of the standard EM algorithm, as described in section 3.2. This requires either changes to standard HMM tools, or the development of a new software, such as described in (Ikbal, Bourlard, Bengio and Weber, 2001).

A second way is to unfold the HMM2 (which, as previously stated, is a kind of HMM mixture) into one large HMM (Levin and Pieraccini, 1993; Kuo and Agazzi, 1993; Samaria, 1994; Eickeler, Müller

and Rigoll, 1999; Weber, Bengio and Bourlard, 2001b), as shown in Figure 4.1. For this implementation, synchronization constraints have to be introduced to ensure that exactly one feature vector is emitted between each two transitions in the primary HMM. This requires (1) additional synchronization states[1] (grey in the figure) and (2) a re-arrangement of the data. Out-of-range synchronization components (modeled exclusively by the synchronization states) are introduced between the original feature vectors. The transitions between primary HMM states correspond to transitions between the synchronization states. Standard EM training algorithms (as presented in Section 1.4.3) can be used to implement this unfolded HMM2, and Viterbi decoding has to be used at the level of both the primary and the secondary HMM.

We did preliminary tests with both of the HMM2 implementations described above. It was found that they yield a similar performance on small problems. For practical reasons, all further experiments reported here used the implementation shown in Figure 4.1, realized with the HTK system (Young et al., 1995).

4.2 Evaluation of Data Representation and Discrimination

In Section 3.2.2 we have presented the assumptions we need to impose on the data in order to model them with an HMM2 system, and in Section 3.3 we discussed their implications in some more detail. In fact, each component is assumed to be independent of all other components, given the HMM state, and a data segment is assumed to be piecewise stationary along both the time and the frequency axes (i.e., a

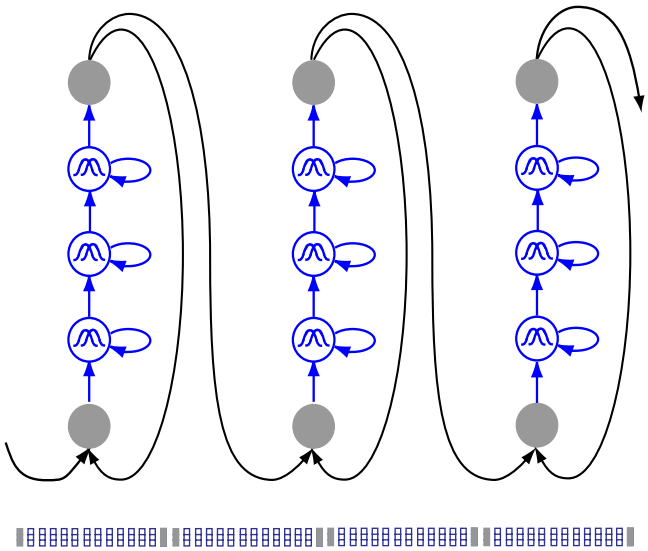


Figure 4.1: HMM2 implementation with synchronization constraints and synchronization subvectors. The HMM2 system is emitting a sequence of (low-dimensional) components, intermitted by synchronization components at regular intervals.

[1]It is sufficient to introduce one synchronization state either at the beginning or at the end of each secondary HMM. However, for the sake of clarity we here choose 2 synchronization states before and after each secondary HMM.

few subsequent components are supposed to have been generated by the same probability density function). As is generally the case for a temporal sequence of speech data, these assumptions may not entirely be satisfied for the sequence of sub-vectors processed by the HMM2 system, which may result in a mismatch between the data and the model's capacity for data representation and discrimination (Weber, Bengio and Bourlard, 2001b). We now investigate whether these assumptions are satisfied, and their significance for the speech data representation in HMM2, as compared to conventional GM-HMM systems.

In Figure 4.2, correlation coefficients of FF2 features are visualized. It can be seen that the data are correlated, especially neighboring components in a feature vector (indicated in the figure by darker colors near the diagonal). Figure 4.3 shows how these correlated data are represented by a GMM and by a secondary HMM. The models are both trained on real FF2 speech data, and their respective parameters are visualized (in the same way as for the toy example in Figure 3.5). In the left part of the figure, it can be seen how the GMM parameters represent the existing data correlation. However, the HMM, shown in the right part, is not able to reproduce an appropriate data distribution. Although there are many suitable methods which orthogonalize data to some extent, completely uncorrelated features have yet to be found in the domain of ASR[2]. If this mismatch between data and modeling capacity cannot be circumvented or compensated for, data representation by HMM2 might remain sub-optimal.

The validity of the stationarity assumption is harder to fully prove or reject. Figure 4.4 shows an example pronunciation of phoneme /ay/. It can be seen that the piecewise stationarity assumption is not entirely satisfied. Nevertheless, it is intuitively (and practically, using a clustering algorithm) possible to segment this representation along the (horizontal) frequency axis in a few quasi-stationary sectors, which could subsequently be represented by the same PDF.

4.2.1 Visual Evaluation of the Frequency Index

In Section 3.4, we have discussed the issue of data discrimination, related to the stationarity assumption and thus to the modeling of a sequence of sub-vectors by one secondary HMM state, and we have introduced a potential way to improve data discrimination by adding additional frequency information to each frequency sub-vector. To evaluate the meaning of such frequency information, an HMM2 system was trained, using secondary feature vectors augmented by a frequency index. In Figure 4.5, the corresponding Gaussian means are shown for different phonemes of the database. The associated variances are also visualized in the figure. While the trained means of the frequency index provide information about the position of the frequency bands modeled by the corresponding states, the variances model the respective bandwidths. It can be seen that these parameters vary across phonemes, and that, for a given phoneme, they may also vary in time. The figure confirms that some general structural information of the phonemes is modeled. However, the structures represented in the figure are not meant to be sufficient for phoneme discrimination, as no supplementary (and generally available) information about other underlying speech features (such as the energy in the different frequency bands) is visualized.

[2]Even the correlation coefficients of (the supposedly decorrelated) MFCC are quite comparable to those of FF2 (shown in Figure 4.2), with the difference of a lower correlation near the diagonal.The issue of the modeling of correlation with HMMs has been discussed in Section 1.6.

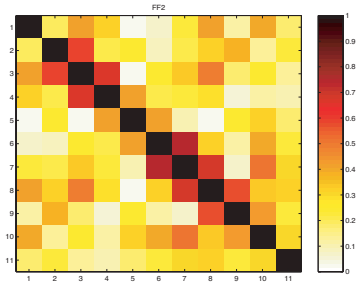


Figure 4.2: Correlation coefficients of FF2 features. Dark colors correspond to high correlation coefficients.

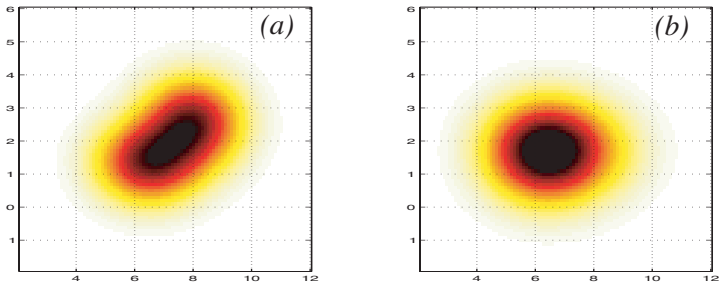


Figure 4.3: Illustration of the modeling power of GMM and Markov model using real FF2 speech data. Figure (a) shows a part of a trained GMM, (b) the equivalent trained Markov model (only two dimensions are displayed). In either case, there are mixtures of 3 Gaussians. While in (a) data correlation becomes obvious, it cannot be seen in (b).

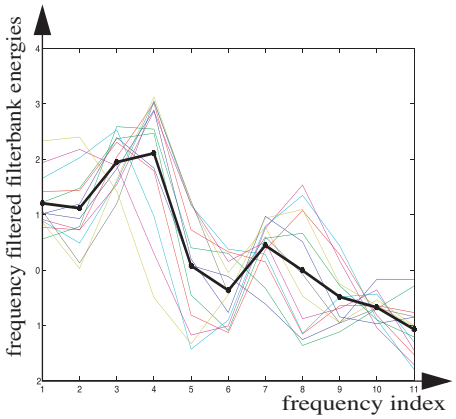


Figure 4.4: Energy spectrum of a pronunciation of phoneme /ay/. Each line in the figure corresponds to one time step, and thus to one feature vector (the thick black line is the mean).

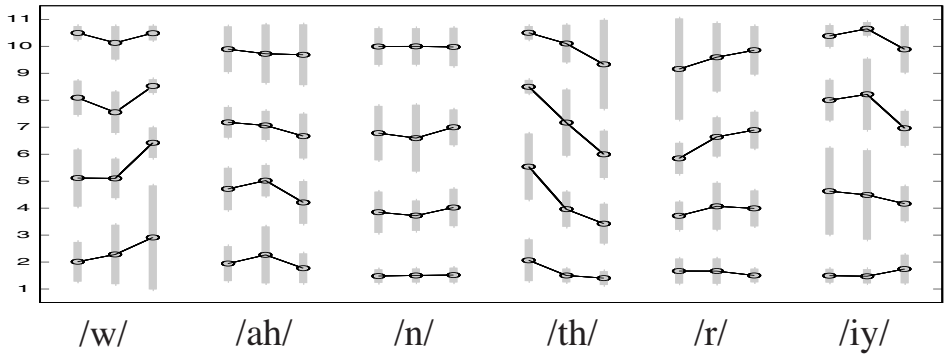


Figure 4.5: Trained HMM2 parameters for different phonemes. In each column, the means of the frequency indices of the 4 secondary HMM states belonging to the same temporal state are visualized. Vertical bars show the respective variances. The 3 columns belonging to a phoneme correspond to the 3 temporal states. It should be noted that these structures are not meant to be sufficient for phoneme discrimination.

4.2.2 Preliminary Evaluation on a Speech Recognition Task

In this section, the HMM2 approach is evaluated on a real speech recognition task. In order to directly compare HMM2 with the conventional GM-HMM system, the topology of the primary HMM was left constant throughout the tests (all phoneme models had 3 temporal states, connected by a strict left-right topology). Only the likelihood estimation in each primary HMM state was changed, realized by GMMs with different numbers of mixtures, or alternatively with different frequency HMM topologies such as described in Section 3.3 and depicted in Figure 3.4. Test were also done using an additional frequency index (FI). Word error rates are shown in Table 4.1.

	Monophones 1 Gaussian	Monophones 10 Gaussians	Triphones 10 Gaussians
GMM	22.2	12.5	6.7
no-loops model	$21.8^{(1)}$	$18.3^{(2)}$	$11.4^{(2)}$
HMM	41.9	$31.6^{(3)}$	$20.5^{(3)}$
HMM/FI	42.2	$27.2^{(3)}$	$15.9^{(3)}$

Table 4.1: Comparison of systems using different models for the local likelihood estimation of the primary HMM: WER on Numbers95. Where applicable, the numbers in superscripts designate the corresponding topology (see Section 3.3).

As in the case of conventional GM-HMM, for each of the tested HMM2 variants performance improves as the model becomes more complex. For both the no-loops model (topologies 1 and 2) and the HMM (topology 3), a mixture of 10 Gaussians performs generally better than a single Gaussian, and triphone models have a superior performance as compared to monophones.

Comparing GMM and the non-looped model, it can be seen that their performance for the case of monophones with a single Gaussian distribution is comparable (22.2% vs. 21.8% WER respectively). This was expected, as the two systems were shown to be theoretically similar (see Section 3.3.1). The slight difference in the results can be attributed to differences in implementation and training algorithm.

Although for both these systems performance increases when adding more Gaussians, the improvement for the case of the non-looped model is inferior compared to the GMM. As was shown in Section

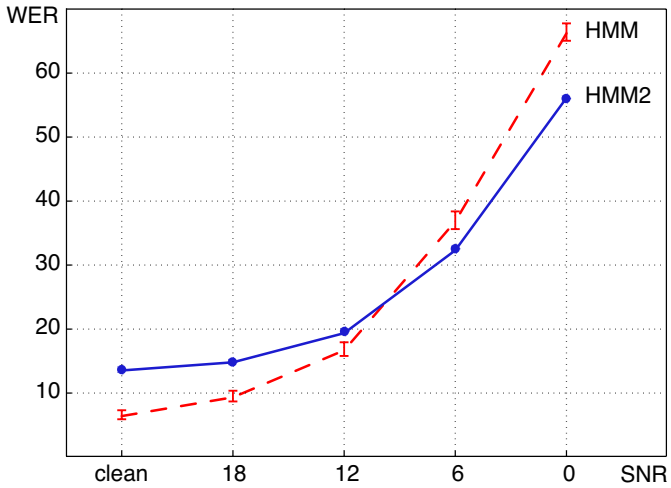


Figure 4.6: HMM vs. HMM2 performance for frequency filtered filterbank features, illustrated by the broken and solid lines respectively, for car noise at different signal-to-noise ratios (SNR). Errorbars for HMM WER show the 95% confidence interval.

3.3.1, a GMM provides better modeling of correlation. Given the correlation in our data (see Figure 4.3), it can be concluded that GMM are indeed the more accurate model for this case. This is confirmed by the results on triphones.

Comparing the no-loops model and the HMM, we encounter yet again a serious performance drop. It could be argued that this loss might be due to the parameter sharing and thus a lower number of parameters in the HMM. However, the HMM with mixtures of 10 Gaussians has about four times as many parameters than the no-loops model with a single Gaussian distribution, and still the latter model performs much better. This might partly be due to the stationarity assumption not being entirely satisfied, and partly due to the loss of discriminability as described in Section 3.4. Indeed, when using an additional frequency index in the feature vectors (HMM/FI), performance increases. This shows that this frequency index effectively improves discriminability. However, performance of the HMM/FI is still inferior to both the GMM and the no-loops model.

Generally, a significant performance drop was observed when using HMM2. Speech recognition accuracy decreased significantly as compared to the conventional GM-HMM system. This result is consistent for different HMM2 implementations (as described above), and holds for all kinds of features tested.

The experiments reported above were all run on clean speech. However, one of the motivations for using HMM2 is its possibly higher robustness in the case of mismatch between training and testing conditions. Therefore, we investigate in the following section the robustness of HMM2 in the case of additive noise.

4.3 Evaluation on Noisy Speech

To realistically compare the performance of the HMM2 system to that of a conventional HMM, tests were performed on both models given the same features (i.e., FF2, as discussed in Section 1.2.5), and using an additional frequency coefficient. Figure 4.6 shows results for one noise condition, with error bars indicating the 95% confidence interval (more results are given in Appendix A). It can be seen that the differences in the performance of these 2 models are statistically significant. While HMM2 is not

competitive with conventional HMMs in clean conditions or noisy speech with a high SNR, for speech heavily degraded by additive noise it outperforms the conventional HMMs. In fact, HMM2 is better able to handle this kind of mismatch between training and testing conditions (as training was done on clean speech only). This was confirmed on all other tested noise conditions. However, the obtained results (for both HMM and HMM2 with FF2 features) are not competitive with the state-of-the-art performance (obtained with conventional HMMs, but employing as features mel-frequency cepstral coefficients, including spectral subtraction and cepstral mean subtraction, see Section 1.2.4). In fact, the performance is limited due to the choice of features in the spectral domain, which were not found to be competitive with cepstral features on noisy data. Although, to further improve HMM2 performance, more research is required first and foremost in the area of the robust extraction of spectral features, these results indicate the potential for applying HMM2 in adverse conditions.

4.4 Conclusion

In this chapter, the application of HMM2 as a decoder for speech recognition was investigated. Firstly, the capacities of HMM2 for data representation and discrimination were evaluated. It can be stated that the additional assumptions imposed on the data through the particular secondary HMM topology investigated here are not always satisfied. This might be one reason for the performance degradations (as compared to standard HMMs) observed in matched training and testing conditions. On the other hand, the improved flexibility of the model might allow for a better performance in unmatched conditions. In fact, it was shown that HMM2 outperformed conventional HMMs for the case of speech degraded by additive noise with low signal-to-noise ratio, using the same features. However, it has to be stated that using state-of-the-art cepstral features in combination with noise-reduction techniques still yields better results in the framework of conventional GM-HMM systems.

Above, the performance of HMM2 was compared to that of conventional HMMs in terms of the word error rate. Additional considerations are the number of parameters, the amount of training data necessary in order to obtain reliable models, and the recognition speed. Due to the parameter sharing done by HMM2, the number of parameters of this model is generally inferior (given the same number of Gaussian mixtures) to that of conventional HMMs (in fact, in the settings tested, the number of parameters of HMM2 was less than half than that of conventional HMMs). As this has a direct influence on the necessary amount of training data, it can be assumed that the HMM2 model could be reliably trained with the training data available. On the other hand, due to the higher number of states in the HMM2 system (in fact, in the settings tested there were about five times the number of states in an HMM2 as compared to the conventional HMM), the HMM2 recognition speed is considerably lower. While recognition speed is becoming a minor issue given the hardware improvements observed during the last years, it might still be a drawback for the case of real-time applications.

Although HMM2 has not yet been found to be competitive with conventional HMMs in terms of the WER, HMM2 might be able to outperform conventional HMMs. When staying with the bottom-up looped HMM2 topology, the use of better spectral features should yield performance improvements. While the focus of our work was on the acoustic modeling part, ongoing research by other researchers sought (and seeks) to improve feature extraction, also in the spectral domain (Macho and Nadeu, 2001). It is likely that spectral features which outperform MFCCs (including noise reduction techniques) in unmatched conditions using standard HMMs would perform even better in the HMM2 framework.

On the other hand, the particular secondary HMM topology investigated here is just one possibility for an HMM2 implementation. Many other topologies could be employed. For instance, one could start

from a topology mimicking the well-working GMMs (such as depicted in the right panel of Figure 3.3), and include additional transitions where they are found to be useful. This approach might better combine the advantages of both the GMM and a secondary HMM for the local likelihood estimation. More generally, an ergodic topology using a large number of states could be employed, which could permit a better modeling of complex structural information and correlation in the speech signal.

As seen above, there are still many possible research directions in the framework of HMM2. However, during the analysis of the mechanisms underlying HMM2 (with a bottom-up looped secondary HMM topology, as described in this chapter), it became obvious that this model implicitly extracts pertinent information about certain structures of the speech signal (as those visualized in Figure 4.5). Consequently, the idea arose to explicitly exploit this information for speech recognition. This is the subject of the following chapter.

Chapter 5

Application of HMM2 as Feature Extractor

In the previous chapters, HMM2 was introduced. It was shown that HMM2 is a special mixture of HMMs, where emission probabilities of a conventional, "primary" HMM are estimated by "secondary" HMMs, one secondary HMM being associated with each state of the primary HMM (see Figure 3.2). In the case of ASR, the primary HMM works along the temporal dimension of speech and emits a time sequence of feature vectors, and, provided that features in the spectral domain are used, the secondary HMM works along the frequency dimension. In fact, each temporal feature vector is assumed to be a sequence of its sub-vectors, where each sub-vector is associated with a particular frequency band (e.g., reflecting its signal energy). If a temporal feature vector is emitted by a certain temporal HMM state, the associated sequence of (frequency) sub-vectors is in fact emitted by the secondary HMM associated with the current temporal HMM state.

Conventional HMM-based speech recognition is done with the Viterbi algorithm, which finds the best (most likely) path through the model, given the model parameters and the data. This method delivers as a by-product the temporal segmentation of the speech signal, i.e. we get to know not only the sequence of the (supposedly pronounced) sub-phone units, phonemes and words, but also the point in time when each of these speech units begins and ends, and therefore implicitly their duration. In the framework of HMM2, and when applying the Viterbi algorithm at the level of both the primary and secondary HMMs, we obtain additionally (from the secondary HMMs) for each temporal feature vector a segmentation in frequency. Like the temporal segmentation, the frequency segmentation is obtained in a principled way, optimizing a maximum likelihood criterion. Therefore, it can be expected that this segmentation also contains meaningful information. In fact, the frequency segmentation of one temporal feature vector might reflect its partition into frequency bands of similar energy, and therefore indicate the position of spectral peaks and valleys. Spectral peaks are related to so-called formants, which may be useful to discriminate between certain speech sounds.

In this chapter, we investigate whether the time and frequency segmentation obtained through HMM2 Viterbi decoding could (directly or in a converted form) be used as (additional) features for a second, conventional ASR system. In the following, we will refer to such features as "HMM2 features".

5.1 Introduction

Let us first examine more closely an HMM2 system processing a speech signal. For conventional speech recognition, the Viterbi algorithm is used to find the sequence of states that best explains the input data, i.e., that has the highest probability (likelihood) of emitting the given data sequence (equa-

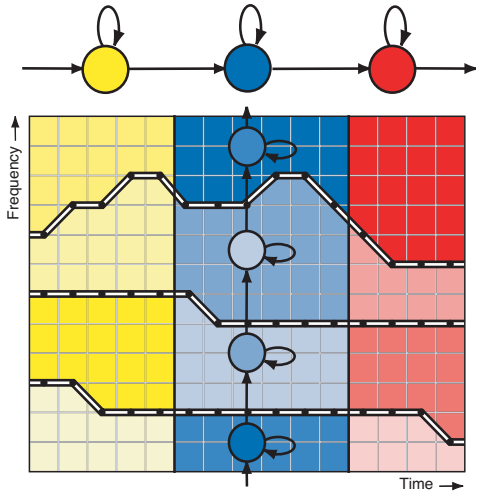


Figure 5.1: Illustration of time and frequency segmentations of a speech signal, as could be produced by HMM2 Viterbi decoding for the example of a 3-state temporal HMM with 4 frequency HMM states each.

tion 3.46). If, at the level of each temporal HMM state, the likelihood of an acoustic feature vector is also estimated using the Viterbi algorithm (equations 3.49 and 3.51), this likelihood is again based on the best sequence of states of the frequency HMM associated with the actual temporal state. Therefore, for the full HMM2 system, we obtain the path through the model that has the highest probability (likelihood) of emitting the observed data sequence, at the level of both the temporal and frequency HMMs. In addition to the sequence of primary and secondary HMM states, we can retain information about where (in time and frequency, respectively) the transitions between these different states take place.

Figure 5.1 shows an example of temporal and frequency segmentations as produced by HMM2. The upper part shows the primary HMM, which segments the speech signal along the temporal (horizontal) dimension. In the lower part, the resulting segmentations are projected onto a (spectogram-like) time-frequency plane (the colors of the columns correspond to the relevant states of the primary HMM). For the middle column, it is illustrated how the secondary HMM associated with the second (blue) primary HMM state segments the corresponding temporal segment along the frequency axis into four different frequency regions (illustrated by different shades). Obviously, the frequency segmentation depends on the state of the temporal HMM. It is clear however that the temporal and frequency segmentations are not obtained sequentially, but in an integrated way during the global Viterbi decoding. While a temporal HMM state typically corresponds to a phoneme or a sub-phone unit, a frequency HMM state corresponds to a particular frequency band. As illustrated in the figure, the positions and the bandwidths of these frequency bands are not fixed a priori, but depend on the data and on the model parameters. In fact, given the bottom-up frequency HMM topology, the Viterbi algorithm can be expected to group (for each temporal feature vector) adjacent frequency components showing similar characteristics into one state, i.e. frequency band. For example, there might be a high energy region covering several frequency components in the lowest frequencies. These components would be modeled by the first frequency HMM state, and the following low energy region would be modeled by the next state. This is demonstrated for the middle (blue) primary HMM state in the figure, where dark shades correspond to high and light shades to low energy regions. However, it can also be seen that, as the signal characteristics change in time, the distribution of high and low energy regions over the secondary HMM states might

be different for each primary HMM state. So, for the right (red) primary HMM state, the energy in the lowest frequencies is comparatively low.

Above, it was discussed how HMM2 can be used to segment the speech signal into regions of similar energy. From these segmentations, new "HMM2 features" can be obtained for a second recognition pass. Firstly, the temporal segmentations could be used to extract features related to duration. Secondly, the segmentations produced by the secondary HMM can be used directly or in a converted form (e.g., as frequency values) as features. Moreover, this frequency segmentation allows the extraction of additional information, such as the average energy of the feature vectors emitted by each secondary HMM state. Additionally, the likelihood that the sequence of feature vectors are emitted by a certain primary or secondary HMM state might contain meaningful information. After having motivated the use of these different HMM2 features, we will describe in more detail the methods used for their calculation.

a) Extraction of features in a principled way

As described in Chapter 3, an HMM2 system can be trained using the Expectation Maximization algorithm (EM), adapting the parameters of the model in such a way that the likelihood of the observations is guaranteed to increase with each training step. Therefore, the final model parameters can be expected to yield a local maximum of the likelihood. Then, during Viterbi decoding, the path through the model that maximizes the likelihood of the data, given the model parameters, is found. Therefore, both training and decoding (i.e. feature extraction) with HMM2 are based on a maximum-likelihood criterion. As discussed previously, Figure 4.5 visualizes trained HMM2 parameters, and some structural information specific to different speech units becomes apparent. Even though HMM2 recognition (and therefore HMM2 feature extraction) is necessarily prone to errors (as in the case of any other decoder), it can be expected that the resulting HMM2 features, obtained by finding the most likely path through the most likely model, carry significant information for speech recognition.

b) Relationship with formant positions

As discussed above, the segmentation between secondary HMM states, produced as a by-product of the Viterbi algorithm, can be interpreted as a separator between regions of different energy levels in the spectrogram (just as the temporal segmentation separates phonetic units). If, e.g., a distinct high energy region is surrounded by low energy along the frequency dimension, it can be assumed to correspond to a formant. Formants are supposed to represent discriminant information, which has been shown to be useful for speech recognition. More details, motivations and results for the use of formants as well as formant-related HMM2 features for ASR can be found in Chapter 6.

c) Duration-related feature

Apart from the frequency segmentation, also the temporal segmentation produced as a by-product of the Viterbi algorithm might also contain useful information. A transition from one primary HMM state to the next might indicate a change in the characteristics of the speech signal, and the time spent in a primary HMM state might be related to parameters such as phoneme duration and speaking rate. The use of duration related features has been investigated and has shown some success, e.g. in (Wang, 1997).

d) Relationship to Tandem System

Recently, the "tandem system" has been proposed (Hermansky, Ellis, and Sharma, 2000), using phoneme emission probabilities as estimated by artificial neural networks (ANN) as features for conven-

tional HMMs. Applying this approach to the framework of HMM2, the secondary HMM state likelihoods could also be used to calculate a special kind of HMM2 features. These likelihoods may also provide some sort of confidence measure of the recognized speech unit, and therefore of the quality of the (other) HMM2 features.

e) Dynamic multi-band

The relationship between HMM2 and the multi-band approach has already been discussed in Chapter 3, and can be extended to the case where HMM2 is used as a feature extractor. In fact, the Viterbi algorithm finds the most likely path through the primary and secondary HMMs. At the level of each temporal feature vector, the associated sub-vectors can thus be assigned to the secondary HMM states that are supposed to have emitted them. Each such secondary HMM state might represent one frequency band, and would therefore define the frequency band's characteristics (e.g., high or low energy). For each temporal feature vector, the HMM2 frequency segmentation is not defined a priori, but depends on the observation. Therefore, the cut-off frequencies and bandwidths of the frequency bands are adaptive. The application of HMM2 directly as a decoder can be seen as an implicit multi-band implementation where the bandwidths of the different frequency bands are adjusted dynamically, depending on the data. When HMM2 is used as a feature extractor, the frequency segmentations can be used to calculate new features, given these dynamic sub-bands. As all of the sub-vectors emitted by the same secondary HMM state can be assumed to show similar characteristics, it can be assumed that they contain redundant information. They could therefore be compressed, e.g. simply through averaging. Together with the information about the (adaptive) positions (and/or bandwidths) of the frequency bands, the compressed sub-band energies can be expected to represent relevant information for ASR. A feature related to the "contents" of a band might be particularly useful if it is considered that high and low energy regions might be located in different sub-bands for different primary HMM states.

5.2 HMM2 features

As already described to some extent in this chapter, the temporal and frequency segmentations delivered as a by-product of the Viterbi algorithm serve as the basis for calculating new features for a second recognition pass, which we refer to as "HMM2 features". In this section, we describe some techniques for calculating HMM2 features.

5.2.1 Time Index

Let us first consider the temporal segmentation. In most applications, the point in time at which a certain speech unit starts or ends is of no value for discrimination. While the duration of a speech unit in comparison to other speech units might give some clues about its identity (e.g. vowels tend to be longer than consonants, and plosives tend to be very short in comparison to other phonemes), there is a non-negligible correlation between duration and other non-discriminant features such as the speaking rate. Therefore, a duration feature might be of limited use for ASR. However, it might be useful to know whether a certain temporal feature vector has a temporal position near the start, the center or the end of a speech unit.

Given a temporal segmentation, we have direct access to the start time t_s and end time t_e of each speech unit. For each time step t with $t_s \leq t \leq t_e$, a "time index" TI can be calculated using the following equation:

$$TI = 1 - \frac{t_e - t}{t_e - t_s} \tag{5.1}$$

where $0 \leq TI \leq 1$. Therefore, the first temporal feature vector which is supposed to be emitted by a certain primary HMM state is attributed a time index of 0, and $TI = 1$ corresponds to the last emitted temporal feature vector. Intermediate feature vectors have indices equally spaced between 0 and 1. Alternatively, a time index can be computed over several primary HMM states, e.g. over all states belonging to a phoneme (t_s would therefore be the time when the first HMM state associated with a certain phoneme is entered, and t_e the time when the last such state is visited for the last time).

In either case, the exactness of this time index is limited and depends on (1) a sufficiently good temporal segmentation, which is likely to be influenced by the HMM2 recognition performance and (2) the sampling rate of the temporal feature vectors (usually, a temporal feature vector is extracted every 10-20ms).

5.2.2 Frequency Index

Unlike the temporal segmentation, the frequency segmentation may directly represent discriminant information. As shown in Figure 5.1, each transition from one secondary HMM state to the next corresponds to a transition between frequency regions, which are represented by the different frequency sub-vectors. As the spectral features represent discrete frequency components, the frequency segmentations obtained from the secondary HMMs are also discrete. Naturally, the coarseness of the features limits the segmentation capability of the HMM2 not only in the temporal, but also in the frequency dimension. Typical spectral feature vectors $y_{t(1,S)}$ consist of about $S = 10 \ldots 26$ coefficients (i.e. frequency components). If the number of states of the secondary HMM is N_i, there are only $S - N_i + 1$ different frequency values at which a transition from a certain state to the next can take place. This is illustrated in Figure 5.2. When adhering to such a small number of frequency sub-vectors (which might be desirable for practical reasons and necessary in order to achieve a good HMM2 recognition performance), it becomes clear that the frequency segmentation can only be very crude. In fact, a straight-forward mapping of the frequency HMM transitions to integer indices can be used, e.g. indicating the frequency sub-vector that came before (or after) a transition from one secondary HMM state to the next. If desired, these "frequency indices" FI can be mapped to frequency values f_{FI}. An example transformation for the case of filterbank coefficients equally spaced on the frequency axis between $0Hz$ and the maximum frequency f_{max}, and where FI is defined as being the number of the first sub-vector that is emitted by a certain frequency HMM state, is given by the equation

$$f_{FI} = (FI - 0.5) \cdot \frac{f_{max}}{F + 1}. \tag{5.2}$$

However, there seems to no advantage in using real frequency values instead of integer indices as HMM2 features. Moreover, a mapping of the indices to precise frequency values (like the above) might be questionable, as the frequency regions used to calculate the sub-vectors are usually overlapping (as is the case of, e.g., the conventionally applied filterbank analysis, demonstrated in Figure 5.2), and there is no unique boundary frequency value separating two adjacent sub-vectors.

As discussed above, when spectral data such as filterbank coefficients are used as features, the segmentation obtained from the secondary HMMs could correspond to the transition between high and low energy regions. High energy regions might be related to formants. Therefore, the segmentations before and after a high energy region could specify the frequency values between which one (or several) for-

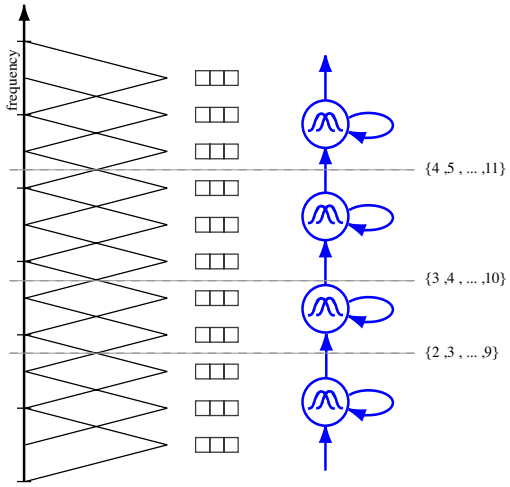


Figure 5.2: Mapping of frequency segmentations to the frequency scale.

mant(s) might be found. Therefore, the frequency segmentations do not correspond to the formant positions themselves. If a closer correspondence to formant values is required, the signal processing could be adapted accordingly (e.g., using frequency-filtered features (as described in Section 1.2.5), as was actually done in most of our experiments), in order for HMM2 to extract spectral maxima and minima. Also, some constraints in order to assure a certain smoothness of the formant tracks might be suitable. Furthermore, a more complex topology of the secondary HMM should be considered. These issues will be addressed in Chapter 6.

5.2.3 Sub-band Energies

The HMM2 frequency segmentations alone do not give any information about the "content" of the resulting frequency bands. As discussed above and shown in Figure 5.1, low and high energy regions might be at different positions, i.e. they might be attributed to different frequency bands. It thus seems appropriate to investigate the use of additional features, such as sub-band energies. A straight-forward implementation of this idea is to take the average or median of the values of all the sub-vectors which are supposed to be emitted by the same frequency HMM state:

$$\overline{x_{t(b)}} = \frac{1}{f_{h(b)} - f_{l(b)} + 1} \sum_{f = f_{l(b)}}^{f_{h(b)}} x_t^f \tag{5.3}$$

where $f_{l(b)}$ and $f_{h(b)}$ determine the low and high cut-off frequencies of sub-band b (resulting from the assignment of frequency components to a certain secondary HMM state by the Viterbi segmentation), and $\overline{x_{t(b)}}$ is the mean of the respective components.

5.3 Practical Issues

5.3.1 Using HMM2 Features in Conventional HMMs

Once the HMM2 features have been calculated, they can be used in a conventional HMM just like any other features. Figure 5.3 shows how HMM2 features are extracted using temporal and frequency segmentations provided in a first (HMM2) recognition pass and then processed by a conventional HMM in

a second recognition pass. Considering the crudeness of the HMM2 features, they cannot be expected to yield competitive recognition performances as compared to more sophisticated and higher-dimensional state-of-the-art features. Nevertheless, their application may increase recognition rates if used in combination with state-of-the-art features. In fact, due to the very different nature of HMM2 features, they may be expected to contain complementary information. Especially in difficult conditions (e.g., noisy speech signal), it is possible that recognition errors made when using HMM2 features are not identical with recognition errors made on conventional features. Therefore, it might be useful to combine both of these features. As has been shown in Chapter 2, a combination can be done e.g. on the feature level, or at the level of the local state likelihoods.

5.3.2 "One Model" Variant

As discussed above and seen in Figure 5.3, an HMM2 system is itself a speech decoder. Consequently, HMM2 recognition is prone to errors (as also seen in Chapter 4). It is clear that, if the HMM2 features are extracted using the wrong model, these features are error-prone too, and are likely to be suboptimal. This suggests that a much simplified HMM2 system, which relies on no or only a limited recognition pass, might be more appropriate. The simplest possibility to realize such a system would be to only consider one model (instead of one model for each phoneme) comprising only one (phoneme-independent) temporal state, with which a secondary HMM with several states is associated. We refer to this HMM2 system as OM ("one model"), in contrast to PDM ("phoneme-dependent model"). The OM system models the emission probability distribution of the data set chosen for training. Typically, the whole training set (regardless of the labeling) is used in order to estimate this distribution. Alternatively, a certain subset (e.g., vowels only) may be used for training. It is important to note that this HMM2 system by itself is not intended for speech recognition, but to model some common properties of the training data. However, the frequency segmentation obtained from a forced alignment of this HMM2 can be expected to contain meaningful information. Apart from the time index, all of the HMM2 features described previously can be extracted.

5.3.3 HMM2 Initialization

Different methods can be used to estimate the initial parameters of an HMM2 system[1]. For instance, assuming that each temporal feature vector is composed of alternating high and low energies, the respective Gaussian means (of the initial single Gaussian distributions) can be initialized with high and low energy values (referred to as HL-initialization). Alternatively, for all temporal feature vectors, a linear segmentation along the frequency axis can be assumed, and the respective means (and possibly variances) of the data can be used as initial Gaussian parameters (referred to as MU-initialization). If the temporal labeling is known, phoneme dependent initialization can be done, which should result in more accurate initial parameter estimates. While using these different initializations did not seem to significantly affect the performance of HMM2 when directly applied as a speech decoder[2], the resulting HMM2 features (and their performance) are significantly influenced, as discussed below[3].

[1] In fact, standard initialization procedures (such as starting from a linear segmentation of the data) cannot be used for the unfolded HMM2 because the explicitly introduced synchronization constraints have to be taken into account.

[2] For this reason, this issue was not discussed in Chapter 4.

[3] Different initialization methods (including formant-dependent techniques) were tested during later work and will be discussed in Section 6.3.4.

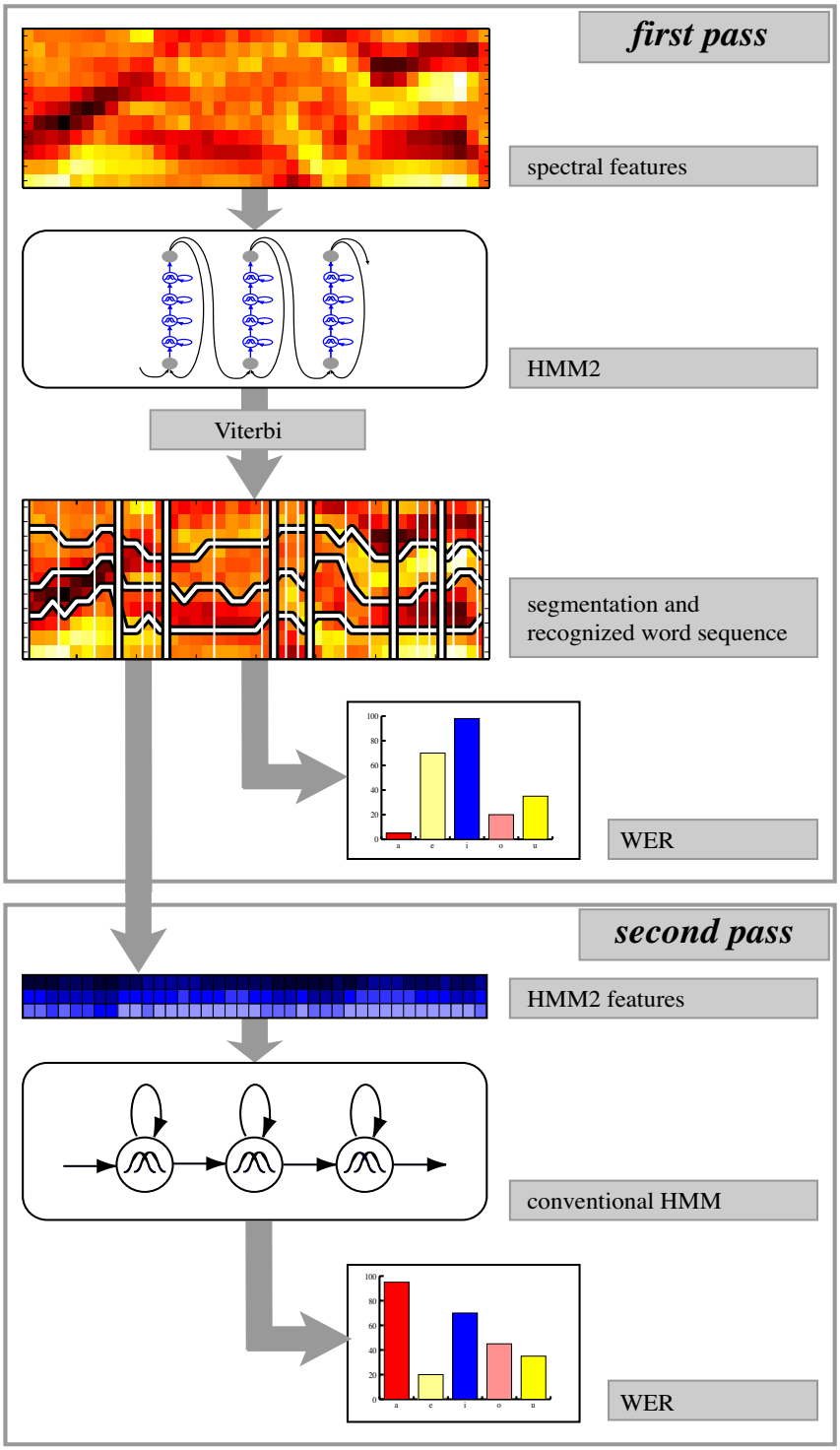


Figure 5.3: HMM2 system in its application as feature extractor. The HMM2 system is used in a first recognition pass (upper part of the figure). From the temporal and frequency segmentations delivered as a by-product from the Viterbi algorithm, HMM2 features can be calculated and used in a conventional HMM in a second recognition pass (lower part of the figure).

5.4 Experiments

Experiments were run using the same experimental setup as described in Section 4.1, i.e. the database was Numbers95, second order frequency-filtered filterbank coefficients were used as features, and HMM2 was implemented in HTK.

5.4.1 Evaluation of Different Kinds of HMM2 Features

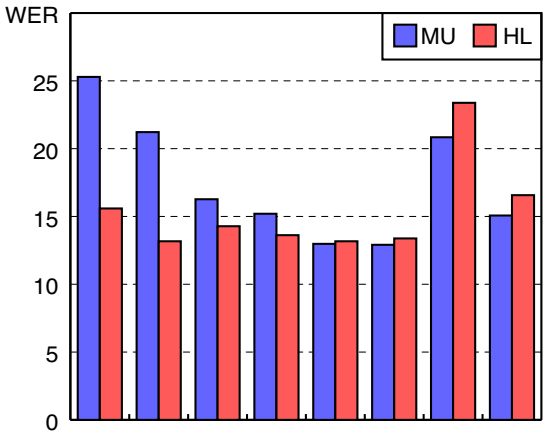
A first set of experiments was carried out using the PDM system to extract different kinds of HMM2 features. In particular, three frequency indices (represented by the number of the component before which transitions between the four subsequent secondary HMM states took place), one time index (calculated according to equation 5.1, where (t_s corresponds to the time when the first of the 3 primary HMM associated with a phoneme was entered, and t_e to the time when the third primary HMM state was visited for the last time), and the average subband energies (calculated using equation 5.3). As an additional feature, the overall energy of the entire temporal feature vector was used.

Different initialization methods (as described in Section 5.3.3) were tested. Figure 5.4 visualizes two test series. The first series was done with a phoneme-dependent initialization of the Gaussian means (referred to as MU in the following). The corresponding results (in terms of word error rates) for the different HMM2 features extracted using this system are displayed on the left of each cluster. The second series was done using an initialization based on the assumption of alternating low and high frequency bands (referred to as LH. In fact, the Gaussian means of the respective FF2 coefficients were simply assigned values of 1 or -1 respectively). The right bar of each cluster shows the results obtained with this system. For each cluster, it is indicated in the table below the bar graph which features were used[4].

Let us first look at the results obtained with the HMM2 features extracted by the MU system (left bar of each cluster). Using only 3 frequency indices as features, WERs of about 25% were achieved. Using additional temporal derivatives of these features (indicated by “xda” in the corresponding field of the table), the error rates were decreased by about 4%. When a time index was used instead of these temporal derivatives, even better results were obtained. Adding the overall energy as an additional feature further reduced the error rate. Finally, the best results were obtained when also the sub-band energies were appended (resulting in a 9-dimensional feature vector). However, using additional first and second order temporal derivatives (which gave a feature vector with 27 dimensions) did not improve recognition. The two clusters on the far right illustrate the results obtained when only using overall and sub-band energies, with and without first and second order temporal derivatives respectively, resulting in rather mediocre error rates.

Looking now at the results of the LH system, it can be seen that it outperforms the MU system for most of the different HMM2 features. In fact, the LH system gave a WER close to the best MU system result even when only 3 frequency indices were used as features. However, it is interesting to note that this tendency is reversed for the case where only energy features are used. This might be explained as follows. For the MU system, the initial Gaussian means of the frequency HMM states were calculated given a segmentation linear in frequency for each temporal feature vector. Therefore, the initial system is tuned toward this linear segmentation, resulting in only minor variations of the FI. The corresponding sub-band energies are therefore calculated on the base of relatively stable sub-bands. On the other hand, although the initialization of the LH system might seem comparatively crude, it was chosen with the

[4]In fact, the recognition error rates of these two HMM2 systems when used directly as a decoder were 13.0% for the MU and 15.5% for the LH system respectively.



frequency index FI	x	xda	x	x	x	xda		
time index TI			x	x	x	xda		
overall energy				x	x	xda	x	xda
sub-band energies					x	xda	x	xda
resulting dimension	3	9	4	5	9	27	5	15

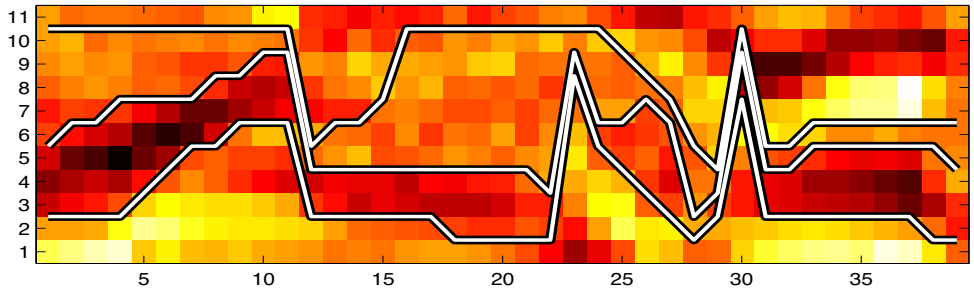
Figure 5.4: Word error rates obtained using different features extracted from a MU and an HL HMM2 system (displayed by the left (blue) and right (red) bar of each cluster respectively). With each cluster of the bar graph in the upper part of the figure, one column of the table below is associated. The features that were used for the respective tests are marked with an "x". The notation "xda" signifies that additional first and second order time derivatives were used. The last row of the table shows the resulting feature dimension for each setting.

aim of separating high from low energy regions. Therefore, the frequency index features is more likely to contain discriminant information such as formant positions. However, if such an LH-segmentation is indeed obtained in a similar way for each phoneme, the corresponding sub-band energies might not provide significant additional discriminant information.

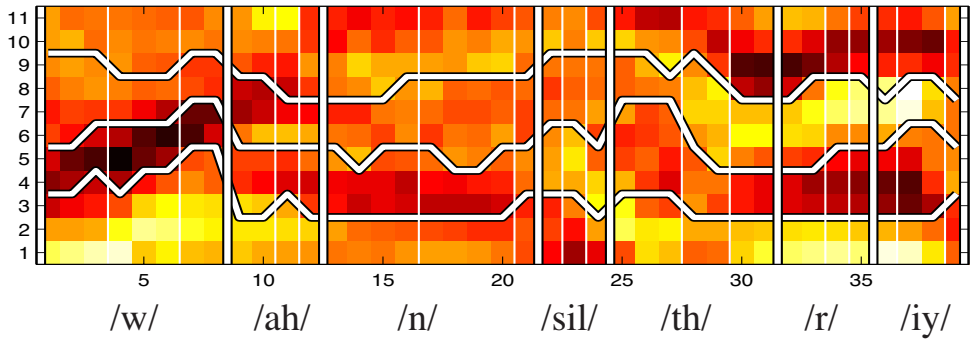
5.4.2 Evaluation of Features From Different HMM2 Systems

In a second set of experiments, the OM and PDM systems were compared. The OM system (featuring just one primary HMM state) was trained on all data regardless of the labeling, as described in Section 5.3.2. In this case, no frequency coefficient (as described in Section 3.4) was appended to the secondary feature vectors[5]. The frequency segmentation for an example speech unit is shown in Figure 5.5a. It can be seen that different secondary HMM states seem to model spectral regions of different energy. In particular, the HMM state modeling the second lowest frequency band constantly emits coefficients of comparatively high energies. E.g., in the beginning of the displayed speech segment, there is a high energy region, whose maximum moves with time from relatively low to relatively high frequencies. The overlaid segmentations (1st and 2nd lines from the bottom of the sub-figure) follow this evolution. This is followed by a rather abrupt change in the speech signal's characteristics, reflected by a sudden transition of all segmentations to different frequencies, and the same state finds again a spectral maximum,

[5]In fact, this system was also tested including a frequency coefficient. For the tested parameter setting, this however resulted in a uniform frequency segmentation, naturally providing no discriminant information.



(a) segmentation obtained from a single secondary HMM



(b) segmentation (and labeling) obtained from a full HMM2

Figure 5.5: Segmentations obtained (on unseen data) from (a) a single secondary HMM and (b) a full HMM2 system. In both figures, the same speech segment is shown in a spectrogram-like manner, and the overlaid horizontal lines correspond to the frequency segmentation. In (b), additional vertical lines show the temporal segmentation obtained from the full HMM2 system, where phoneme boundaries are displayed as thick lines, and transitions between temporal states of the same phonemes as thin ones.

now in relatively low frequencies, until the next change of signal characteristics. In fact, in the case of less distinct or absent formants (as for the case of unvoiced phonemes), irregularities and discontinuities can be observed.

Figure 5.5b shows the segmentation obtained from a PDM system, including the frequency coefficient in the secondary feature vector. The segmentation is smoother, but high and low energy regions are not necessarily modelled by equivalent secondary HMM states (e.g., the third secondary HMM state may model a high energy region for one phoneme and a low energy region for another). Nevertheless, a certain structure of the speech signal becomes apparent from the segmentation.

The features extracted by different OM and PDM systems were tested in a second HMM recognition pass. For the OM system, word error rates on clean speech were between 35% and 45%, depending on the particular parameter settings. As seen above, with the PDM system, error rates are in the range of about 15% to 25% when using the 3-dimensional frequency index features. Although for the case of PDM the meaning of the frequency segmentation is not necessarily consistent for all phonemes, this system clearly outperforms the simplified OM system.

5.4.3 Combination with MFCCs

Given the baseline performance of about 5.7% word error rate obtained on this database when using MFCCs as features, it is obvious that HMM2 features are not competitive. Also, none of the HMM2 fea-

ture combinations tested in Section 5.4.1 were able to outperform the respective HMM2 system when applied as a decoder. In spite of the motivations for using HMM2 features outlined above, their utility might be questioned given these experimental results. On the other hand, we have shown previously how the robustness of a state-of-the-art ASR system based on MFCCs could be improved by using additional features in a second feature stream. Indeed, for the case of HMM2 features it can be expected that they present complementary information to that contained in MFCCs, which could be exploited in such a multi-stream system. Consequently, HMM2 features from both the OM and the PDM system were also tested in combination with noise-robust MFCCs (already including spectral subtraction (SS) and cepstral mean subtraction, denoted MFCC-SS in the following). The noise conditions presented in Section 2.1.1 were used.

Table 5.1 gives an example for the performance of the OM system in the case of additive Noisex factory noise (more results are given in Appendix B). Word error rates are compared to those of MFCC-SS. Although the HMM2 features (i.e., 3-dimensional frequency index features) yield very high error rates, positive results were obtained when using them in (feature) combination with MFCC-SS. Similar results were obtained on car and lynx noise, and it can be stated that the improvement (as compared to the MFCC-SS baseline) is significant with more than 95% confidence.

SNR	MFCC-SS (baseline)	HMM2 features (OM)	MFCC-SS + HMM2 features
clean	5.7	43.2	5.6
18	7.4	42.3	7.3
12	11.9	49.8	11.4
6	23.0	62.2	**21.4**
0	48.6	76.4	**46.6**

Table 5.1: Word error rates using MFCC-SS, HMM2 features and their combination for clean speech and speech degraded by additive factory noise at different SNRs.

Similar tests were carried out with HMM2 features obtained from PDM systems. Although their performance on clean speech is generally much higher than that of OM in clean conditions, recognition is severely impaired in the case of noise. In contrast to the OM case, the HMM2 features obtained from PDM are error-prone, as they are possibly extracted using the wrong HMM2 phoneme model. Naturally, the lower the quality of the signal, the more recognition errors are made by the HMM2 system, and the lower is the quality of the resulting HMM2 features. Therefore, the HMM2 features obtained from the PDM system are possibly less robust than those obtained from the OM system.

This was confirmed by our experiments combining HMM2 features with MFCC-SS. In fact, performance improvements were generally less important using HMM2 features extracted by a PDM than by an OM system. An exception to this is the case where the combination was done not at the feature level, but at the local likelihood level (corresponding to feature vs. likelihood combination as discussed in 2.2.1). Comparing the best PDM with the best OM-based feature based system (using feature and likelihood combination with MFCC-SS respectively), performance differences were marginal.

To summarize, it was found that HMM2 features extracted using both the PDM and the OM system can achieve a higher robustness in combination with MFCC-SS than MFCC-SS features alone.

5.5 Conclusion

In this chapter, we have shown how an HMM2 system can be used as a feature extractor, and what kind of features it can provide. The different HMM2 features, as well as their combination with MFCC-SS, were tested for clean speech and under different additive noise conditions. While the HMM2 features alone are not competitive with MFCC-SS, an improved noise robustness was observed when both types of features were used together in a multi-stream approach.

However, it has to be noted that an HMM2 system is rather complex as compared to conventional feature extractors. This is reflected by a multitude of design options, resulting in a large number of possible hyper-parameters and parameters, of which only very few could be tested. Moreover, an HMM2 system is first and foremost a decoder (as investigated in Chapter 4). If a full HMM2 decoder (i.e., an HMM2 system featuring phoneme-dependent models) is used for feature extraction, the resulting HMM2 features are obtained using the most likely phoneme model- which is not necessarily the right one. Consequently, HMM2 features are a priori impaired by the fact that the HMM2 decoder itself makes recognition errors. This can be avoided when using only one model for all data, where no recognition takes place, and only the optimal path through the (single) secondary HMM is found and converted into features. However, it has been shown that the error-prone, phoneme-dependent HMM2 features (from a full HMM2 system) perform significantly better than features obtained from this simplified HMM2 feature extractor.

These results suggest that on the one hand, the modeling capabilities of the OM might not be sufficient for extracting competitive HMM2 features. On the other hand, the more sophisticated PDM system goes along with a higher confusability, which might be contra-productive as well. This suggests that one way to improve the HMM2 feature extractor could be to use a "compromise" between the OM and PDM system. For instance, one may use a few models which represent broader speech categories than phonemes (such as vowels, fricatives, plosives, etc.). Such a system is likely to be less prone to recognition errors than PDM, which may lead to more efficient and more robust HMM2 features.

In summary, when using a full HMM2 system for feature extraction, classification errors accumulate over the first pass (made by the HMM2 system itself) and the second pass (using the resulting HMM2 features in a conventional HMM). Therefore, classification performance using HMM2 features can not be expected to be higher than that of either of these two system components. One may then ask what advantage is gained by extracting HMM2 features? Firstly, HMM2 features might provide complementary information to that of features conventionally used for ASR. These two feature streams can be easily combined in conventional HMMs. We have shown that speech recognition robustness is improved when using HMM2 features in addition to MFCCs. Secondly, one might be interested in the structural information provided by HMM2 features, as this information might give clues about formant positions. In the following chapter, the relation of HMM2 features to formants is investigated in more detail.

Chapter 6

Formant-related HMM2 Features for ASR[1]

In the previous chapter, it was shown how HMM2 can be used as a feature extractor. In fact, the segmentations produced by an HMM2 system as a by-product of Viterbi decoding can be used to generate "HMM2 features", which can be exploited as (additional) features in a second HMM recognition pass. It was empirically demonstrated that the Viterbi frequency segmentation may separate regions of low energy from regions of high energy. Therefore, it can be expected that these frequency segmentations may be related to formant frequencies. In this chapter, we take a closer look at those frequency segmentations and their relation to formant values. After having briefly discussed formant extraction techniques and the use of formants in ASR, we will give some detail about the HMM2 formant extractor. Then, we will focus on an empirical investigation, using what experts labeled as "formants" as features for speech recognition. We investigate their capacity for vowel classification, and compare their performance to that of MFCCs. Then, the obtained results serve as a reference for the comparison with automatically extracted formant-related features, namely "Robust formants" and HMM2 features.

6.1 Formants in ASR

Formants may be defined as the resonance frequencies of the vocal tract. In a spectrogram, they can usually be distinguished by the presence of high energy (i.e., spectral peaks) in the concerned frequency bands. In the context of speech production, formants are explained using a tube model of the vocal tract. The shape of this tube defines its frequency selectivity, and when the shape is changed, different sounds are produced. It is clear that, as no two speakers have the same vocal tract shape, formant frequencies are not only influenced by what is being said, but are also quite speaker-dependent. Nevertheless, it has repeatedly been shown that formant frequencies can well be used to discriminate between different vowels, suggesting that formant frequencies contain more information about different speech sounds as compared to speaker-dependent characteristics.

One often referenced study of vowel acoustics was done over 50 years ago by Peterson and Barney (1952). As reported e.g. in (Rabiner & Schafer, 1978; Hillenbrand et al., 1995), Peterson and Barney used a spectrograph to measure the formant frequencies of vowels, which had been pronounced by 76

[1]This chapter is partly based on work done with Febe de Wet, Loe Boves and Bert Cranen from the University of Nijmegen, The Netherlands. A lot of the reported experiments were done in tight collaboration. However, while work on HMM2 features was done exclusively by the author of this thesis, the "Robust Formants" parts are contributions of the above-mentioned colleagues.

male, female and children speakers in /h-V-d/ syllables, and which were perceived to be equivalent. If the frequencies of the second formant (F2) are plotted against those of the first formant (F1), it can be seen that, even though there is a lot of variability within each vowel and some overlap between different vowels, vowels can be separated quite nicely in this F1/F2 plane. Although a later study by Hillenbrand et al. (1995) claims that the overlap of different vowels is actually more important than the measurements by Peterson and Barney indicate, it was confirmed in the same study that good results can be achieved when using formant frequencies as features for automatic vowel classification.

However, there are only a few state-of-the-art speech recognition systems which actually use formants or formant-like features. One of the reasons lies certainly in the difficulty of automatically estimating them. In the following, we will give a short description of formant extraction methods and their application to ASR.

6.1.1 Formant Extraction

In this section, we will give a short overview of some formant extraction techniques. Many common formant extractors are based on linear prediction analysis (McCandless, 1974), performing a frame-by-frame computation of the roots of a linear predictor polynomial (Atal and Hanauer, 1971; Talkin, 1987; Lee et al., 1999), or searching for maxima in the spectral envelop (also referred to as "peak picking") (Schafer and Rabiner, 1970; Laprie and Berger, 1994). Alternative approaches include analysis by synthesis (Olive, 1971), possibly based on digital resonators (Welling and Ney, 1998). Furthermore, methods to find the spectral peaks can be based on banks of bandpass filters (Padmanabhan, 2000), energy gravity centroids (de Mori et al., 2000), or mixtures of Gaussians (Zolfaghari and Robinson, 1996; Stuttle and Gales, 2001). However, most of these methods suffer from several disadvantages, such as (1) the number of formants (or spectral peaks) found might vary from frame to frame and (2) the formant tracks are often not smooth (as one would expect given the physiological constraints of speech production). As a result, these methods extract at best formant candidates. One of the reasons for these problems lies in the fact that formants are often not well-defined, or ambiguous. In fact, there is generally no one-to-one relation between the spectral maxima of an arbitrary speech signal and its representation in terms of formants, and there may be more or fewer prominent maxima, depending e.g. on the spectral characteristics of the source signal.

One technique overcoming the above problems is the "Robust Formants" algorithm (Willems, 1986) which will be described in more detail below. An alternative way is to impose continuity constraints on the formant tracks (Schafer and Rabiner, 1970; McCandless, 1974). This kind of formant trajectory optimization can be done for example using dynamic programming (Ney, 1983; Talkin, 1987). A related approach uses HMMs for formant tracking (Kopec, 1986). On the other hand, one can make use of the phonetic labeling (Lee et al., 1999). If the phonetic labeling is known, the a priori distribution of "formant targets" can be used in order to chose the most likely formants from a number of formant candidates. For example, this can be done using Viterbi search in a second formant tracking step (Acero, 1999; Huang, Acero and Hon, 2001). While using information about the phonetic labeling may improve the accuracy in formant estimation, for obvious reasons, techniques relying on this kind of a priori information can not be applied for the case of speech recognition in general and the study presented in this chapter in particular.

An alternative way, which can be used in ASR but still does not completely disregard information about the phonetic labeling, is to delay the selection of formant tracks until after a phonetic search has been carried out (Schmid and Barnard, 1995), or to directly combine formant tracking with phoneme recognition (Hasegawa-Johnson, 1996). This is also motivated by the assumption that the "analysis of

formants separately from hypotheses about what is being said will always be prone to errors" (Holmes, 2000).

Above, we briefly introduced a variety of formant extraction techniques. We have discussed that formant extraction may be improved if the phonetic labeling is known. However, this is generally not the case in the context of ASR, where, to the contrary, one may use information about formant trajectories in order to find the phonetic labeling. Before discussing the use of formant features in ASR, we will give a short introduction to the "Robust Formants" algorithm as one example for conventional formant extraction techniques, producing formant values suitable as features for ASR.

6.1.2 The "Robust Formants" Algorithm

The "Robust Formants" (RF) algorithm (Willems, 1986) guarantees to provide a fixed number of formants at each time step, and ensures a certain smoothness of the resulting formant tracks. This algorithm was initially designed for speech coding and synthesis applications. It uses the Split Levinson Algorithm (SLA) to determine a fixed number of spectral maxima for each speech frame. Instead of directly applying a root solving procedure to a standard LPC polynomial to obtain the frequency positions of the spectral maxima, a so-called singular predictor polynomial is constructed from which the zeros are determined in an iterative procedure. All the zeros of this singular predictor polynomial lie on the unit circle, with the result that the number of maxima that are found is guaranteed to be half the LPC order under all circumstances. The maxima that are located in this manner are referred to as "formants" found by the RF algorithm.

After the frequency position of the RF formants have been established, their corresponding bandwidths are chosen from a pre-defined table such that the resulting all-pole filter minimizes the error between the predicted data and the input. The frequencies at which the zeros of the classical root solving procedure occur are close to the unit circle (i.e., as long as the true formants have small bandwidth values). This property ensures that the most important formants are properly represented.

In contrast to standard root solving of the LPC polynomial (or searching for maxima in the spectral envelop derived from LPC coefficients), the RF algorithm finds a fixed number of "formants" for each speech frame. This makes the RF algorithm particularly suitable for the goal of using the extracted formants as features for ASR, because the algorithms that are used in ASR are generally designed to deal with feature vectors of a fixed length.

6.1.3 Formant Features for ASR

As already discussed above, formants are useful for discrimination between certain speech sounds, and there are numerous attempts of incorporating them in ASR systems. For instance, in (Welling and Ney, 1998), a recognition system based solely on formant contours is presented. Comparing these features to a mel-cepstrum representation (with the same number of parameters), recognition results on clean speech were reported to be only slightly better for the latter case. Another system uses formant trajectories in combination with formant bandwidth, pitch and segment duration, achieving a comparable performance to a cepstral-based system on a vowel/semi-vowel segment classification task (Schmid, 1996).

However, it is argued in (Holmes, Holmes and Garner 1997) that formant frequencies cannot discriminate between speech sounds for which the main differences are unrelated to formants. Therefore, formants were used not instead of, but in addition to features such as MFCCs, leading to a better recognition performance than obtained on MFCCs alone. Other systems where additional formant related fea-

tures were found to improve recognition performance in certain conditions include (Padmanabhan, 2000), (de Wet et al., 2000), and (Stuttle and Gales, 2001).

6.1.4 HMMs and HMM2 as Formant Extractor

As mentioned above, the idea of using HMMs for formant extraction is not new (Kopec, 1986, Hasegawa-Johnson, 1996). In particular, Kopec's formant tracker HMM emits a sequence of temporal feature vectors, just like conventional HMMs used for ASR. However, in this case, the HMM states correspond to possible formant values (and not to, e.g., phonemes). Continuity constraints are implemented in the transition probabilities of the HMM (i.e., there are high transition probabilities to states representing close formant positions). For training, a database with hand marked formant-tracks is required. In (Hasegawa-Johnson, 1996), an extension to this system is proposed, where multivariate-state HMMs are used to simultaneously transcribe phonemes and formants. Another HMM-based extractor of formant-related structures, which moreover may also be used for phonetic classification, is the HMM2 system, as explained below.

In Section 3.1.2, we have motivated HMM2 (amongst other arguments) by its ability to implicitly extract structural information of the speech signal, possibly corresponding to formant regions. In the same line of thought, it was discussed in Chapter 5 that HMM2 might be capable of separating high from low energy regions. The fact that formants can be expected to be located in the extracted high energy regions opens up the perspective of using HMM2 as a formant tracker. HMM2 also offers the advantage that continuity constraints can easily be incorporated. Moreover, for every time frame, a fixed number of "formants" is found, which facilitates their application as features in ASR. Although the interpretation of the HMM2 frequency segmentation as formant-like regions may not always be fully justified (as seen later), this application is additionally motivated by HMM2 being a tool which can integrate a speech decoder and a formant tracker in a unique model, as discussed above (Holmes, 2000).

6.2 HMM2 Formant Extractor

In the previous chapter, the HMM2 feature extractor was explained in detail. Let us recall that for each temporal feature vector, it is determined between which sub-vectors a transition from one frequency HMM state to the next takes place. In fact, the number of the first sub-vector being emitted by a new frequency HMM state (i.e., just after a transition in the frequency HMM took place), can be retained as an index, which can directly be mapped onto a frequency value. Depending on the features used, these frequencies might correspond to spectral peaks and spectral valleys or transitions between them (as will be discussed in more detail below).

6.2.1 Preliminary Study

The purpose of this section is to investigate if, under ideal conditions, a frequency HMM is indeed capable of extracting structural information related to spectral peaks and valleys, and thus possibly to formants. Therefore, a preliminary study was carried out, in which a much simplified HMM2 topology was used. In fact, the temporal HMM consisted of just one state, and the associated frequency HMM was a 4-state bottom-up HMM, corresponding to the OM system discussed in Section 5.3.2. However, training was done with data (from the Numbers95 training set) labeled to belong to one vowel (/iy/), and testing was done on the same data.

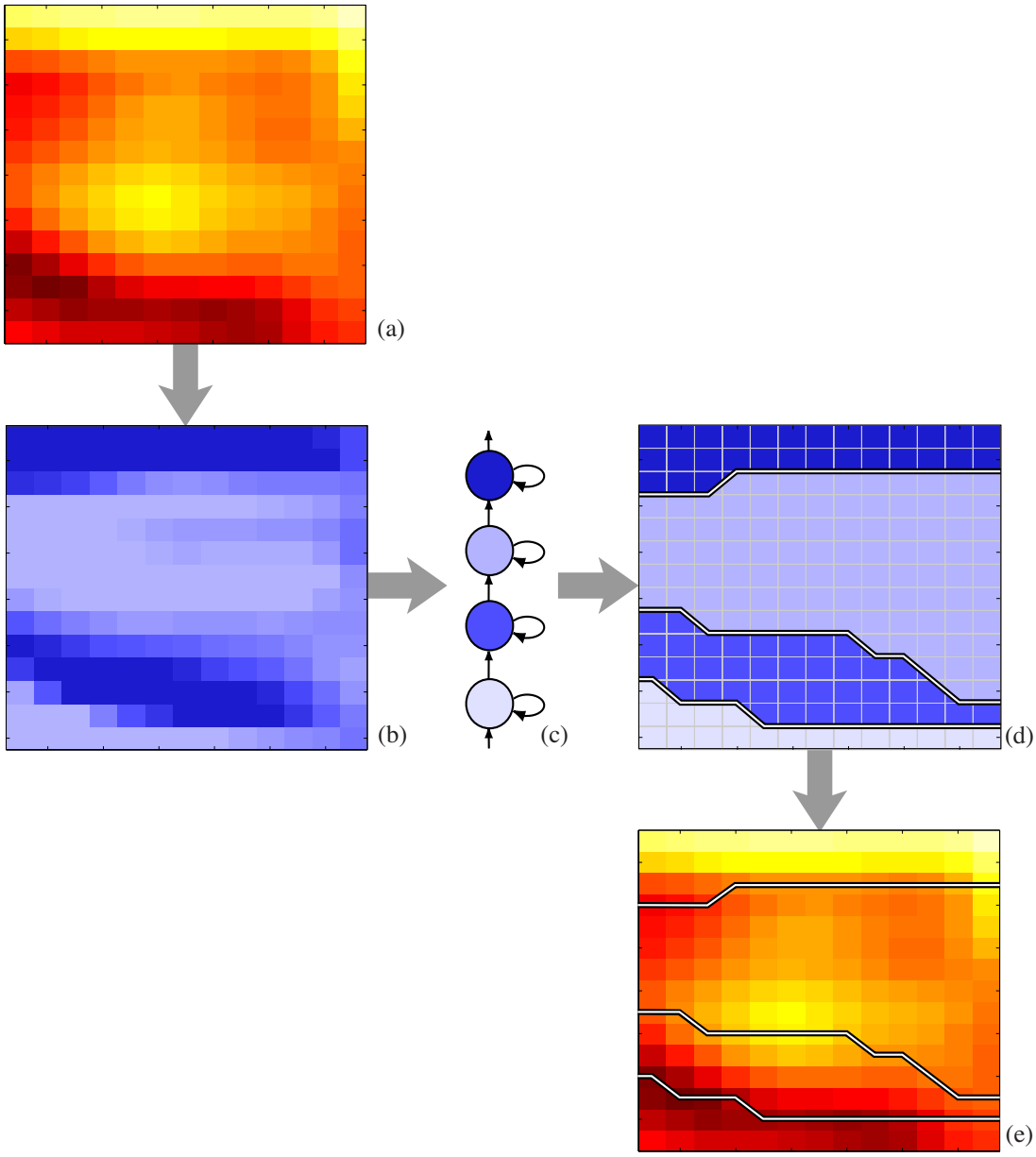


Figure 6.1: Features and segmentations for an example of phoneme /iy/. Sub-figure (a) shows a spectrogram-like representation of log Rasta PLP features, and (b) their respective first order frequency derivatives. In (c), the topology of the frequency HMM is shown, and in (d), the frequency segmentation obtained from a forced alignments of the data in (b) given this HMM is shown in the time/frequency plane. In (e), the projection of this segmentation onto the original features is visualized.

For obvious reasons, features in the spectral domain were chosen. Each temporal feature vector consisted of 14 frequency derivatives of log Rasta PLP features (Hermansky et al., 1991), obtained simply by taking the difference of each two adjacent log Rasta PLP coefficients, using first order frequency filtering (as described in Section 1.2.5). Using those delta features has two advantages for our application. On the one hand, the frequency differencing implicitly introduces a kind of normalization (Nadeau, 1999). On the other hand, we can take advantage of the fact that these features are some kind of derivatives, positive values corresponding to increasing energies (along the frequency dimension), and nega-

tive values to decreasing energies respectively. As the boundary between regions of increasing and decreasing energy is obviously a spectral maximum, a formant could be expected at this frequency location. If one frequency HMM state was to model positive changes in energy, and the surrounding states were to model negative changes, then the transitions between these states would fall onto spectral peaks and valleys of the original, undifferentiated features, and every other frequency segmentation might fall onto (or near a) formant.

An example of the original (15-dimensional) features can be seen in Figure 6.1(a), and their frequency derivatives are visualized in Figure 6.1(b). While in Figure 6.1(a) dark/light regions correspond to high/low spectral energies, in Figure 6.1(b), dark/light regions correspond to negative/positive changes in energy. In Figure 6.1(d), the segmentation as performed by the frequency HMM Figure 6.1(c) is visualized. It can be seen that states 2 and 4 tend to model high values, whereas states 1 and 3 model low values. In particular, the segmentation of state 3 seems to model the "dark region" moving in time and frequency. In Figure 6.1(e), the projection of this segmentation onto the original log Rasta PLP features (i.e., not the frequency deltas) can be seen. In this case, a transition of the frequency HMM corresponds to a coefficient (rather than to the transition between coefficients, as for the frequency derivatives). These transitions follow approximately the maximum and minimum energy regions, but taking account of constraints imposed by the topology and parameterization of the frequency HMM. Although a possible correspondence to formants has not been proven, it can be noted that the segmentations between the first and second as well as between the third and forth state approximately correspond to spectral peaks.

However, it needs to be stressed that this is a very preliminary test, which was done under ideal conditions. For instance, data of only one phoneme (the identity of which was known) was used for training and testing. The features used might not be optimal when applied in GM-HMMs for application in ASR. Moreover, using only 14 coefficients severely limits the resolution of the resulting frequency segmentation. Finally, the topology of the frequency HMM was chosen such as to obtain an optimal match with the data to be modeled, which could intuitively and invariably be divided into high/low regions along frequency.

In this section we have demonstrated what could ideally be expected from an HMM2 system towards the goal of extracting formant-like features. However, our main goal of extracting formant-like features is their use for speech recognition. In the following, we will first show that "true" formant features, obtained from hand-labeled formant tracks, are indeed useful for a vowel classification task, and compare their performance to that of MFCCs. Furthermore, another method to automatically extract formants is investigated. All these different features are then compared to HMM2 features in terms of their classification performance.

6.3 AEV Database, Experimental Setup and Baseline System

In the preceding section, it was argued that HMM2 could be used to extract features which are related to formant frequencies. The purpose of this section is to evaluate this assumption by comparing HMM2 features to hand-labeled formants (HLF) and other automatically extracted formant-like features. However, databases featuring hand-labeled formant tracks are rare and difficult to obtain. As the hand-labeling of formant frequencies requires a considerable human effort, these databases are typically rather small, and often cover only vowels. The research reported in the following is thus necessarily constrained by the limitations imposed by the used database (which was the only database containing hand-

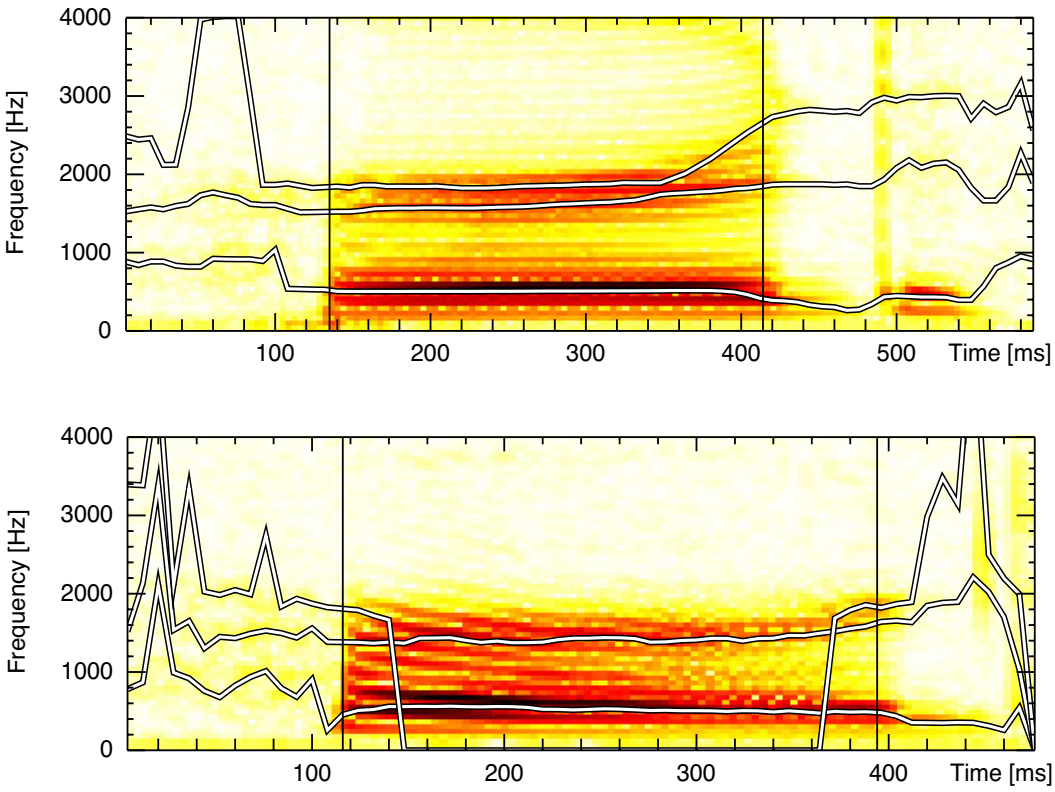


Figure 6.2: Example spectrogram and formant tracks (F1,F2 and F3) of two pronunciations of the phoneme /er/ (pronounced within the word "heard"), as provided with the AEV database. Only the frequency band from 0-4000Hz is shown. The vertical black lines show the part which was labeled as the vowel part, according to the segmentation provided with the database. It can be seen that the formant tracks corresponding to the leading /h/ and trailing /d/ are very irregular. In the lower figure, a merger of F2 and F3 occurred, and the upper frequency slot was thus set to zero.

labeled formant tracks available to us). In particular, instead of speech recognition, a vowel classification task is investigated (using, however, ASR technology).

6.3.1 Database of American English Vowels

The speech material that was used for most of the experiments presented Chapter 6 is a subset of the database of "American English Vowels" described in Hillenbrand et al., 1995. This database contains recordings of the 12 vowels produced by 45 men, 48 women and 46 children. The vowels are embedded in /h/-V-/d/ syllables, i.e., there is an /h/ proceeding and a /d/ following the vowel. The speech signals are studio quality and were digitized at 16 kHz. Various acoustic measurements were made for each token in the database, including vowel duration, vowel steady state times[2], formant tracks and fundamental frequency tracks. In what follows, the focus will be on the formant tracks, since these values were used as features in our classification experiments.

[2]Vowel steady state was defined by Peterson and Barney as "... following the influence of the /h/ and preceding the influence of the /d/, during which a practically steady state is reached" (Peterson & Barney, 1952).

To obtain the formant tracks, candidate formant peaks were first extracted from the speech data by means of a 14th order LPC analysis. These values were subsequently edited by trained speech pathologists and/or phoneticians. In addition to the LPC peaks overlaid on a gray-scale spectrogram, labelers were also provided with individual LPC or Fourier slices where necessary. The labelers were allowed to repeat the LPC analysis with different parameters and to hand edit the formant tracks. The formant tracks were only hand edited between the start and end times of the vowels, i.e. the formants corresponding to the leading /h/ and trailing /d/ of the /h/-V-/d/ syllables were not manually labeled. Where unresolvable formant mergers occurred, the higher of the two formant slots affected by the merger was set to zero. Two examples of the data provided in the AEV database are shown in Figure 6.2. The lines which are overlaid onto the spectrogram correspond to the three lowest formants F1, F2, and F3.

Hillenbrand et al. (1995) showed that the vowel classes can be separated reasonably well (in comparison with human performance) by applying a quadratic discriminant analysis (QDA) on the values of the first three formants measured at a number of pre-defined times in the vowel.

6.3.2 General Experimental Setup

In all the experiments reported in this section, a subset of the AEV database was used, consisting of the 12 vowels pronounced by 45 male and 45 female speakers. Only the vowel parts of these utterances were taken into consideration. This allows a direct comparison between the hand-labeled formants and all other features, because the formant tracks of the leading /h/s and trailing /d/s were not hand-edited.

In comparison with the databases that are typically used in ASR experiments, the AEV database is quite small. Given this limitation, a 3-fold cross-validation was used for the classification experiments. Each experiment consisted of a number of independent tests, in which the models were trained on two subsets of the data, and tested on the third one. Moreover, all tests were performed in two conditions, i.e. gender-independent and gender-dependent. The gender-independent data sets were defined as three non-overlapping train/test sets, each containing the vowel data of 60(train)/30(test) speakers, with an equal number of males and females in each set. For the gender-dependent data, three independent train/test sets were defined for males and females, respectively. Each train/test set consisted of 30(train)/15(test) speakers. For the gender-independent data sets, the classification results reported below correspond to the mean value of the three independent tests. The gender-dependent results were obtained by averaging the classification results of the six independent experiments (three male and three female).

Feature extraction was done on speech data downsampled to 8kHz[3]. All acoustic analyses adhered to the same time resolution used in (Hillenbrand et al., 1995), i.e. the frame rate was set to one frame per 8ms. For each of the feature sets described below and for each of the mixed/male/female cross-validation sets defined above, a three state HMM was trained for each vowel using the EM training algorithm implemented in HTK. Each state consisted of a mixture of 10 continuous density Gaussian distributions.

Using these basic definitions, the following feature were tested:

- MFCC, as state-of-the-art ASR features,
- FF2, employed as the basic features for the HMM2 feature extractor,

[3]Although this might not be optimal for the frequency filtered filterbank features which are the base for HMM2 feature extraction, this option was chosen as it is consistent with the other methods described in this section and moreover with the experiments on the Numbers95 database reported in previous chapters.

- HLF: hand-labeled formants F1, F2 and F3, as provided with the AEV database (described in Section 6.3.1),
- RF: robust formants, i.e. automatically extracted formant tracks with the method described in Section 6.1.2, and
- HMM2 features.

Before evaluating the vowel classification performance of formant-related features (i.e., HLF, RF, and HMM2 features) in Section 6.4, we first give some details on the results obtained with MFCCs, and describe the HMM2 system setup (including results on FF2).

6.3.3 MFCC Baseline Results

In this section, a standard baseline system for the present vowel classification task is established, using standard ASR techniques. In particular, Mel-frequency cepstral coefficients (MFCC) are used as features, as they are employed in many conventional ASR systems. As stated before, a quite common configuration is to use 13 MFCCs (including energy, denoted in the following as MFCC-13) and their first and second order time derivatives (denoted D and A respectively). However, the feature dimension of such a system is much larger than that of systems using exclusively formant features (13 vs. 3 feature vector components). For that reason, test were also done on 3-dimensional MFCCs, using the first three coefficients[4] (and no energy, denoted as MFCC-3). Table 6.1 gives an overview of the baseline system results for 13- and 3-dimensional MFCCs, as well as for the same features with additional time derivatives.

Feature Type	dim	Gender-independent	Gender-dependent
MFCC-13	13	88.1	89.4
MFCC-13 + D	26	92.3	92.1
MFCC-13 + D + A	39	90.6	90.9
MFCC-3	3	68.7	78.2
MFCC-3 + D	6	77.6	81.2
MFCC-3 + D + A	9	78.2	79.6

Table 6.1: Classification rates of MFCC features when used in conventional HMMs.

While this table is meant as a baseline and thus as reference for the following sections, there are several points worthwhile noting when comparing the different features. As expected, MFCC-13 performed significantly better than MFCC-3. For most cases, best performances were obtained when the first order time derivatives were included, and performance dropped when additional second order time derivatives were used. Especially for the low-dimensional features, the gender-dependent tests are better than the gender-independent ones. MFCC-13 + D outperformed all other settings for both gender-independent and gender-dependent tests.

[4]However, it should be noted that the choice of using the first three MFCCs might not be the optimal one. Although these coefficients contain a lot of discriminant information, performance could certainly be improved if performing a feature transformation (e.g., PCA or LDA, such as described in Section 1.2.3) over the entire feature vectors in order to calculate new 3-dimensional features.

6.3.4 HMM2 System Setup and Design Choices

In this section, we will give some details about the FF2 features which were employed for the HMM2 system, as well as the HMM2 system setup. As discussed before, the choice of FF2 features was motivated by the fact that they are rather decorrelated features in the spectral domain whose baseline performance is comparable to that of other widely used state-of-the-art features such as mel frequency cepstral coefficients (MFCC). In contrast, the Rasta PLP spectra (not to be confused with Rasta PLP cepstra) used in Section 6.2.1 generally achieve lower recognition rates. Moreover, in a previous study, FF2 features outperformed all other features when used for HMM2 (in terms of recognition rates obtained from HMM2 features in the second pass). The particular kind of FF2 features used here is based on fourteen filterbank coefficients, equally spaced on the Mel scale, extracted every 8ms over 16ms long Hamming windows. These filterbank coefficients were then used to compute 12 second order frequency filtered filterbank coefficients (FF2) as described in Section 1.2.5.

For the sake of completeness, let us mention the results obtained when using FF2 features in conventional HMMs. For the experiments reported here, similar settings as for the MFCC baseline system described in Section 6.3.3 were used. Table 6.2 gives an overview of the performance of FF2 features[5]. It can be seen that recognition rates of FF2 features are in the same order as those obtained with MFCC.

Feature Type	dim	Gender Independent	Gender Dependent
FF2 + D	24	90.8	91.8
FF2 + D + A	36	92.4	92.8

Table 6.2: Classification rates of FF2 features when used in conventional HMMs.

However, in contrast to the case of MFCCs, using additional second order time derivatives shows slightly better results. For this reason, the 36-dimensional FF2 + DA features were used as features for the HMM2 system. An additional advantage of using those derivatives is that they may increase the smoothness of the final HMM2 feature tracks (in addition to the effects of the frequency coefficient (as discussed in Section 3.4). Together with their first and second order time derivatives plus an additional frequency coefficient, the FF2 coefficients form a sequence of 12 4-dimensional sub-vectors.

The frequency mapping of FF2 features is visualized in Figure 6.3. In the left part of the figure, 14 conventionally used triangular frequency filters, equally spaced on the mel scale, are plotted on the frequency axis. The approximate cut-off frequencies of these filters are also shown. From the obtained filterbank coefficients, second order frequency filtered filterbank coefficients are calculated through differencing, as shown in the middle part of the figure. On the right, a vector of these FF2, together with the frequency range which contributed to the calculation of each of them, is displayed. Due to the large frequency overlap between two adjacent FF2 coefficients, it is probably not appropriate to define an exact frequency value as the transition frequency (corresponding to the HMM2 frequency segmentation). However, if such a value was required, it seems appropriate to choose the mean value (given the Mel scale) of the frequencies contributing to the adjacent FF2 coefficients, as displayed on the far right.

[5]Preliminary tests, using conventional HMMs, were also done on first order FF coefficients (FF1). For both FF1 and FF2, tests were also done with an additional energy coefficient. Results obtained from all those tests were not significantly different.

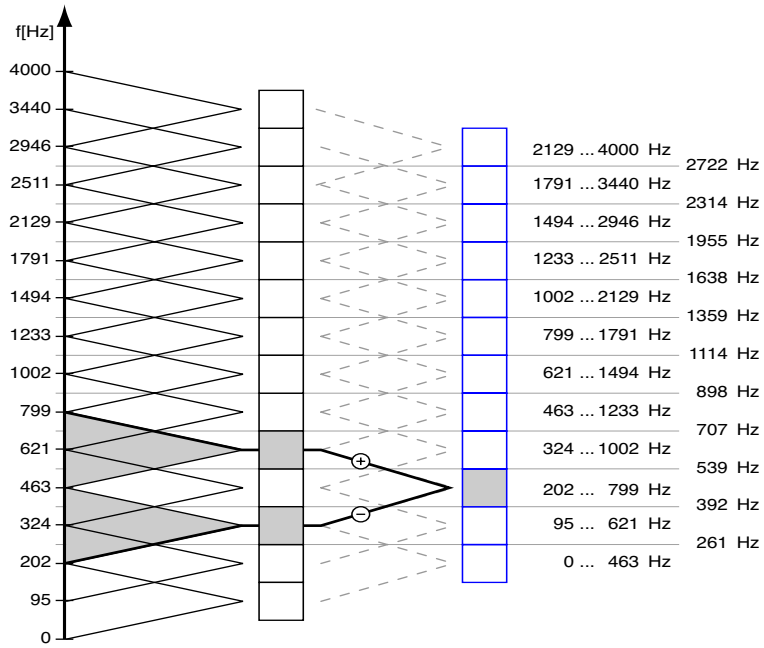


Figure 6.3: Frequency mapping of second-order frequency filtered filterbank features (FF2) as used in HMM2.

Although it seems more appropriate to simply use frequency indices (FI, as discussed in Section 5.2.2), these frequency values give an idea about an approximate segmentation frequencies.

Figure 6.4 shows an example of the features in a spectrogram-like display, where the time evolution is shown on the horizontal and the frequency evolution on the vertical axis respectively. Each square corresponds to one coefficient, where the color intensity indicates its value. In the left part, the speech data is visualized using 14 filterbank coefficients, while in the right part, their frequency derivatives (12 FF2 features) are shown. The corresponding hand-labeled formant tracks of the first three formants (F1, F2, and F3) are overlaid as white lines onto both sub-figures. As expected, for the filterbanks, the formant tracks follow spectral peaks, while they separate high and low energy regions for the case of FF2 features. However, as F2 and F3 are very close, the resolution of the features does not permit a distinction between these two formant tracks.

As in the previous chapters, all training and testing was done with HTK, and the HMM2 system was realized as a large, unfolded HMM, made possible with synchronization constraints as described in Section 4.1.2. As already explained previously, the design of HMM2 systems can vary substantially, depending, e.g., on the task and on the data to model. As before, we here chose a strict "left-right" topology for the temporal HMM (such as typically used for HMMs applied to ASR) and an equivalent "bottom-up" topology for the frequency HMM. With this topology, it can be assured that for each temporal feature vector, 3 segmentation values are found under all circumstances. As stated before, a certain smoothness of these 3 feature tracks is obtained through the temporal derivatives and the frequency coefficient, which are all part of the frequency sub-vectors.

The sequences of 3-dimensional HMM2 features (only consisting of the frequency indices, FI, as described in Section 5.2.2) are then used as features for a conventional HMM in a second recognition pass. As in the baseline experiments, this HMM had 3 states and mixtures of 10 Gaussians were used to model the emission distribution in each HMM state.

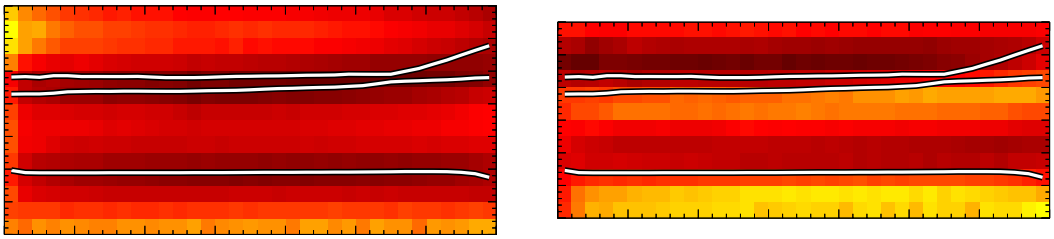


Figure 6.4: Features displayed in spectrogram format. Time evolution is displayed on the horizontal and frequency resolution on the vertical axis. Each square corresponds to a coefficient, where the color intensity indicates different values. In the left panel, 14 mel-scaled filterbank coefficients are shown, which were used as basis in order to extract 12 FF2 features, as displayed on the right. Hand-labeled formant tracks are projected onto both sub-figures.

However, there are some restrictions imposed by the choice of features and model topology, limiting a priori the quality (i.e., exactness) of the HMM2 features. Particularly, HMM2 features are rather crude. As was demonstrated in Section 5.2.2, there are only very few possible frequency segmentation values for any transition to a new state in the frequency HMM. Moreover, in Figure 6.4 it becomes apparent that the resolution of FF2 features does not permit to distinguish between two formants which are close in frequency. In spite of these limitations, 12 FF features can be expected to provide some relevant structural information, which could be extracted by an HMM2 system.

Above, we have discussed the features employed for HMM2 and the choice of HMM2 topology. As discussed before, there is a multitude of other design choices. In Section 5.3, some design options concerning the model and the initialization were already mentioned. These as well as some additional design choices are investigated below.

a) OM and PDM systems

In Section 5.3.2, we have already introduced the OM variant, featuring only one temporal state which is trained on all the training data. Both the OM system and the PDM (i.e., the full HMM2, featuring 3 temporal states per phoneme) system are also considered in the experiments reported in this chapter.

b) Frequency Coefficient

A further HMM2 design decision concerns the use of a frequency coefficient as additional component of the frequency sub-vectors. It has been shown that using this frequency information improves discrimination between the different phonemes. However, the impact of the frequency coefficient is different depending on it being treated (1) as an additional feature component (feature combination, FC) or (2) as a second feature stream, where the likelihoods of the two streams are locally combined for each time step (likelihood combination, LC). In the latter case, additional parameters are the stream weights. In fact, it is possible to use the stream weights for regulating the smoothness of the HMM2 features. The higher the weight of the frequency coefficient stream, the greater is its impact as compared to the other, "real" features, and the more the frequency segmentation is constrained. In its extreme, the frequency segmentation would be constant throughout the data, and consequently the HMM2 features would be meaningless. On the other hand, if the frequency coefficients' weight is too low, the segmentation tends to be very irregular, with sudden transitions, peaks, etc. Setting an appropriate stream weight might assure the desired smoothness of the HMM2 feature track.

c) Initialization

As already discussed in Section 5.3.3, the initialization of the HMM2 models can be done in different ways. For instance, a linear segmentation along the (Mel) frequency axis can be assumed for each feature vector. In the experiments reported here, there are four secondary HMM states. While a higher number of secondary HMM states might be advantageous towards a more formant-like feature extraction, the restriction of using only four secondary HMM states was motivated by the relatively low resolution of the data. Each of the secondary HMM states is initialized using three frequency sub-vectors assigned to it according to the linear segmentation. Another option is to assume an alternation of low (L) and high (H) energy bands of the FF features, and initialize the HMM2 with expected values for these lows and highs, thereby forcing an HLHL or LHLH segmentation along the frequency axis. Alternatively, as formant frequencies are provided with the AEV database, these can be used in order to obtain an initial non-linear frequency segmentation (FMT). E.g., for each phoneme, the parameters of the frequency indices can be initialized to the corresponding average formant values.

d) Training and Testing

For both OM and PDM, training can be done using the EM algorithm. For the OM model, a forced alignment can be used in order to extract HMM2 features for all data. In contrast, for PDM, HMM2 feature vectors can be obtained in two different ways, depending on whether or not the labeling is known. If the labeling is known, forced alignment (FA) can be used to align the speech data to the corresponding HMM2 model and extract the frequency segmentation. Alternatively, and imperatively if the labeling is not known, a Viterbi recognition (VR), using all phoneme-dependent HMM2 models can be employed. In that case, the segmentation finally extracted by the HMM2 system corresponds to the segmentation produced by the HMM2 phoneme model which has the highest probability of emitting the given data sequence. As discussed before, the features extracted by VR suffer from the fact that the HMM2 system makes recognition errors, resulting in sub-optimal HMM2 features, i.e. features extracted by the "wrong" HMM2 phoneme model. In spite the VR HMM2 features being error-prone, it might be advantageous to use them even for the train data because of the fact that it can be seen as training in a kind of "noise", or "matched" conditions.

Different combinations of FA an VR can be used for the train/test data. As in a classification application the labeling is generally not known for the test data, we focused on FA/VR and VR/VR. However, an FA/FA system could be realized if the labeling was known also for the test data, e.g. if the task is not classification, but to extract HMM2 features for a labeled speech segment- or, as in our case, to estimate a theoretical upper limit of the system performance, assuming HMM2 recognition is perfect.

6.4 Experimental Results on Formant-related Features

6.4.1 Evaluation of Formant Tracks for ASR

The purpose of the experiment described in this section was to evaluate the classification performance of true formant features using state-of-the-art speech recognition methods. In (Hillenbrand et al., 1995), it was demonstrated that first three formant frequencies, extracted at defined points in time of the vowel duration (e.g., "steady state", or 20%, 50%, and 80% of vowel duration) represented discriminant information and achieved good classification results when Quadratic Discriminant Analysis (QDA) was applied. However, in a speech recognition task, the segmentation is generally not known, which makes

it impossible to extract features at such defined points in time. Therefore, methods like QDA can generally not be applied.

Here, we investigate the performance of formant features using HMMs. We used three HMM states, where the emission distributions were modeled by mixtures of 10 Gaussians. The entire formant tracks were used (i.e., the values of F1, F2, and F3, in 8ms intervals over the whole vowel duration). As described in Section 6.3.1, where mergers occurred in the hand-labeled formants, the frequency of the higher formant was set to zero. However, it seems more judicious to resolve these mergers, e.g. by replacing the zeros in the higher formant slots by the frequency values in the lower ones, therefore using two equal values. Experiments have been run on the original formant tracks (including zeros) and on the new formant tracks with resolved mergers, and no significant performance differences were observed. It can be assumed that in the case of the original tracks, one Gaussian mixture component takes care of the zeros, and the remaining 9 components are still sufficient to model the other data. Moreover, in keeping with what has become standard practice in ASR, the formant frequencies can be mel-scaled. Again, experiments on the formant tracks on the linear as compared to the mel scale did not show significant differences.

As mel-scaled formant tracks with resolved mergers are more consistent with what might be expected from the methods used for automatic extraction of formant related features discussed below, the results reported here are based on this variant of HLF features. Results are shown in Table 6.3. Vowel classification rates of 83.2% for the gender-independent and 85.9% for the gender-dependent experiments were achieved. In both cases, the performance improved when first order time derivatives were added. Although these results are not competitive with what Hillenbrand et al. (1995) reported on QDA[6], it can be stated that good recognition rates can be achieved when using ASR technology for formant features.

Feature Type	dim	Gender Independent	Gender Dependent
HLF	3	83.2	85.9
HLF + D	6	87.2	89.6
HLF + DA	9	86.5	89.2

Table 6.3: Classification rates of hand-labeled formants.

In comparison with MFCC-13, HLF features achieved significantly lower classification rates, for both the gender-independent and the gender-dependent cases. However, the performance of MFCC-3 is far below that of both MFCC-13 and HLF[7].

[6] In contrast, Hillenbrand et al. (1995) reported 91.8% on gender-independent tests using QDA for 3 samples at 20%, 50% and 80% of vowel duration. This difference might partly be attributed to differences in the used data sets (using the whole AEV database instead of a subset, as well as larger train sets), but the main reason lies probably in the discriminant training method for the QDA classifier, in contrast to conventional EM training for the HMM. These issues are discussed in more detail in de Wet et al. (2002), where additional results using an LDA (linear discriminant analysis) classifier are presented.

[7] In fact, Figure 6.7 summarizes these results and the results reported further below. Moreover, confidence intervals are given.

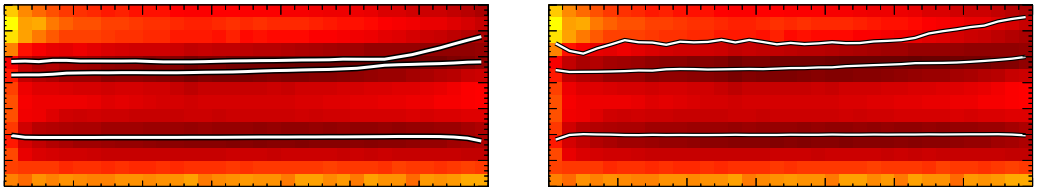


Figure 6.5: Hand-labeled formant tracks (left panel) and "Robust Formants" (right panel) for one example of phoneme /er/, overlaid onto a spectrogram-like representation of FF2 features.

6.4.2 Evaluation of Robust Formants

In this section, we evaluate the performance of one kind of automatically extracted formant features: Robust Formants, as briefly introduced in Section 6.1.1. We will first give some specifications about the experimental setup directly concerned with the RF extraction, followed by classification results and a visual evaluation.

As mentioned before, the AEV data was downsampled to 8 kHz. It is usually assumed that there are four vocal tract resonances in this frequency band. However, the data in (Hillenbrand et al, 1995) show that F4 could not be found in 15.6% of the vowels. The scope of this study is therefore limited to F1, F2, and F3. Moreover, in the AEV database the mean value (taken over all the relevant data) of F4 is 3.536 kHz for males and 4.159 kHz for females. Thus, it is clear that an automatic formant extraction procedure applied to the AEV corpus must be able to deal with a potential discrepancy between the "true" number of formants in the signal and the requirement that only the first three formants must be returned. For the RF extractor, the simplest way to cope with this requirement that only three formants should be found is to use a 6th order LPC analysis. However, the accuracy of the LPC analysis is bound to suffer if a 6th order analysis is used to analyze spectra with four maxima. In these cases an 8th order LPC would seem more appropriate, although it would introduce the need to select three RFs from the set of four.

Given these constraints, there are a number of choices that can be made concerning the calculation of the RFs. We considered two of these: (1) calculate three RF features per frame (RF3), and (2) calculate four RF features per frame and use only the first three (3RF4). These two sets of RF features were subsequently calculated every 8 ms over 16 ms Hamming windowed segments. The output of the two procedures was evaluated by means of a frame-to-frame comparison with the hand-labeled formants, in terms of their Mahalanobis distance. It was shown (de Wet et al., 2002) that the RF features are closer to the HLF features if the order of the analysis is chosen according to the gender-specific properties of the true formants. If there is a mismatch between the number of spectral peaks the algorithm tries to model and the number of spectral maxima that actually occur in the data, the distance between the automatically derived data and the hand-labeled data increases. Thus, the distance between RFs and the hand-labeled formants decreases if the order of the analysis corresponds to the inherent signal structure. Therefore, in the gender-dependent experiments, the RF3 and 3RF4 features will be used for the female and male data, respectively. However, as for the gender-independent tests the gender is obviously assumed to be unknown, the RF3 features were used in this case, as they yielded the smallest Mahalanobis distance for the mixed data set.

a) Analysis of Results on RF Features

Table 6.4 shows classification rates for Robust Formant features, for both the gender-independent and the gender-dependent cases. Moreover, results including first order time derivatives are shown. It can be seen that, for both the gender-independent and the gender-dependent systems, performance improves significantly when these augmented (6-dimensional) feature vectors are used, in comparison to using the 3 RF values alone. This indicates that, as in the case of HLF, additional information on spectral change patterns contributes to the discrimination between different speech sounds.

Feature Type	dim	Gender Independent	Gender Dependent
RF	3	76.1	86.3
RF + D	6	84.1	90.5

Table 6.4: Classification rates of Robust Formants.

As for the case of HLF (Table 6.3), the gender-dependent RF system works better than the gender-independent one. In fact, while for the gender-dependent case the classification rates of RF features are comparable to those obtained with the gender-dependent HLF system, the performance of RF are significantly lower for the gender-independent case. This may be attributed to the fact that, in contrast to the HLF case, different RF features are used for these two conditions. As described above, for the gender-dependent tests different features were used for female and male data (RF3 and 3RF4 respectively). This kind of optimization is possible because it is assumed that the gender of the speaker is known. However, for gender-independent tests this kind of a priori knowledge is usually not available, and therefore the same set of (sub-optimal) features (RF3) had to be used throughout. Consequently, in addition to a more difficult modeling of the gender-independent data due to a larger overlap between different phoneme classes (from which gender-independent systems using either HLF or RF features suffer), the RF features themselves are inherently worse for the gender-independent case than for the gender-dependent case.

b) Visual comparison

Figure 6.5 compares Robust Formant tracks (obtained using the gender-dependent set-up, shown in the right panel) with HLF (left panel). It can be seen that the automatically extracted formant tracks are fairly similar to the hand-labeled ones. The example suggests that the spectrum of this vowel contains multiple peaks in the F2-F3 region, and that the automatic RF procedure has generally preferred a peak at a higher frequency than the human labeler. This effect might be explained by the tendency of the RF algorithm shift formants which are close (or merged) away from each other. In addition, the RF features exhibit more frame-to-frame variance than their hand-labeled counterparts, especially for F3. The dip in the F3 track at the vowel onset may be due to the fact that there the “true” formant (frequency peak) was comparatively strong so that the RF procedure could find it, despite its close proximity to F2. Although the RF feature tracks differ form the HLF tracks, the results in Table 6.3 and Table 6.4 suggest that the differences do not seem to significantly affect classification rates. This might be explained by the fact that the differences between RF and HLF may be consistent (i.e., in our example, F2 and F3 would be close for all pronunciations of this vowel, and therefore F3 would be systematically overestimated). This indicates that, to obtain competitive classification performances, it might not be necessary to extract feature tracks which resemble as close as possible to true formants. It might be more important to extract consistent (formant-related) features than true formants.

6.4.3 Evaluation of HMM2 Features

In order to evaluate HMM2 features, all of the design, initialization and training/testing options introduced in Section 6.3.4, as well as combinations of them, have been tested. Most important results using the best HMM2 system are reported below, while more detailed results are given in Appendix C. The systems were compared in terms of the vowel classification performance obtained from the respective HMM2 features. However, no tests were done in order to evaluate the resemblance of HMM2 features to formant tracks, or in terms of formant-related constraints such as the smoothness of the feature track. Best classification results were obtained with 12 phoneme-dependent HMM2 models, with a 3-state, left-right topology in the time domain and a 4-state bottom-up topology in the frequency domain. Frequency coefficients used as a second feature stream gave best results for OM, but the 12 PDMs finally used generally showed a higher performance when including the frequency coefficients as additional feature components in the frequency sub-vectors (FC). For the gender-independent HMM2 models, the LHLH initialization worked best. However, for the gender-dependent models the FMT initialization seemed to be more advantageous, allowing to directly consider the gender-related formants (characterized by generally higher formant frequencies for female than for male speakers) in the initial model parameters. The HMM2 features that are used for training are best obtained by means of forced alignment while those that are used for testing should obviously be obtained from a free recognition (FA/VR). However, for most cases the VR/VR system showed a comparable performance. Naturally, the FA/FA system outperformed both FA/VR and VR/VR, but as this setting is unrealistic, this result has only theoretical value as an upper limit of HMM2 feature performance for the used parameter settings.

In summary, the HMM2 system which was used for the experiments reported in the following section had the following setup:

- 12 phoneme-dependent HMM2 (with 3 primary states, composed of 4 secondary states each), which were realized with HTK as large, unfolded HMMs (as described in Section 4.1.2),
- processing, at every time step, a sequence of 12 4-dimensional frequency sub-vectors comprising a second-order frequency filtered filterbank (FF2) coefficient and its first and second order time derivatives as well as an integer frequency coefficient in a single feature stream,
- using the assumption of subsequent low-high-low-high bands of FF2 coefficients for initialization of the gender-independent models, and using an initialization of the frequency index based on phoneme-dependent average formant values for the gender-dependent models,
- and using forced alignment to obtain HMM2 features of the train set, and Viterbi recognition to obtain HMM2 features of the test set.

Tables 6.5 gives an overview of the performance of HMM2 features, obtained from the HMM2 systems using the settings as described above. It can be seen that these features compare well to the RF features in Table 6.4. Like RF features, HMM2 features are competitive to HLF features for the case of the gender-dependent systems, while the performance is significantly inferior to that of HLF for both automatically extracted formant related features in the case the gender-independent models.

Feature Type (Initialization method)	dim	Gender Independent	Gender Dependent
HMM2 (FMT)	3	71.2	**87.2**
HMM2 (LHLH)	3	**77.0**	83.0

Table 6.5: Classification rates of HMM2 features (second pass) obtained with FA/VR.

Tests were also run on HMM2 features including first order time derivatives, where significant performance drops were observed. In contrast, for all other features tested, performance improved when their first order time derivatives were included. The poor performance of HMM2 feature derivatives can be explained by the very crude nature of the HMM2 features. As seen above, they consist of only a few integer values, and HMM2 feature tracks are typically constant over relatively long time intervals, and otherwise present sudden jumps (or even oscillations). After a more general analysis of the errors made by HMM2 features, HMM2 feature tracks will be visually evaluated, demonstrating these problems.

a) Analysis of Results on HMM2 Features

As explained before, HMM2 features (from a full HMM2 system) are prone to errors, as they are extracted in a first recognition pass (on the base of FF2 features processed in an HMM2 system). In a second recognition pass, these HMM2 features are used in a conventional HMM, which is error-prone as well. As a result, employing this kind of two-pass system means that recognition errors can be made at two different levels. In the following, the contribution of errors of the two passes is investigated.

Table 6.6 shows classification rates achieved by HMM2 in the first pass (i.e. when HMM2 is applied as a decoder, not as a feature extractor). Comparing results with those obtained when using conventional HMMs (given the same features, see Table 6.2), they are significantly worse for the gender-independent data. Results on the gender-dependent data do not differ significantly. However, the errors made by the HMM2 system mean that a significant proportion of the HMM2 features has been extracted using the wrong model. As a result of this first recognition pass, HMM2 features are therefore inherently error-prone.

Feature Type (Initialization method)	dim	Gender Independent	Gender Dependent
FF2 (FMT)	12*4	86.5	**91.7**
FF2 (LHLH)	12*4	**88.2**	91.8

Table 6.6: Classification rates of FF2 features when used in an HMM2 system in the first pass (using different initializations).

To determine the contribution of the second recognition pass to the error rates, HMM2 features were extracted under the assumption that HMM2 recognition is perfect. Given the correct phoneme model, forced alignments was applied for training as well as for testing. Using the resulting "ideal" HMM2 features in the second recognition pass, we get an idea about the theoretical upper limit of the HMM2 features extracted by the two selected models. Results are shown in Table 6.7. Again, performance on the

Feature Type (Initialization method)	dim	Gender Independent	Gender Dependent
HMM2(FMT)	3	77.9	**93.6**
HMM2(LHLH)	3	**81.6**	89.0

Table 6.7: Classification rates using "ideal" HMM2 features (obtained from FA/FA) in the second recognition pass.

gender-independent sets is much worse than that on the gender-dependent ones. In fact, comparing these features to HLF (Table 6.3), these "ideal" HMM2 features are not as good as their HLF counterparts for the gender-independent sets, but significantly better for the gender-dependent sets. It might be argued that using "ideal" HMM2 features signifies doing recognition given a priori the knowledge of

what has been said. However, this situation (unrealistic for the case of ASR) is in fact quite similar to what happens when formant tracks are hand-labeled by professionals. Usually, the class of the speech segment is known to the labeler, and the labeler will (consciously or not) use a priori information about expected formant positions during the labeling process. This basically means that, when using hand-labeled formant tracks as features (as well as "ideal" HMM2 features) for phoneme classification, knowledge is used that normally is not available.

Comparing the results in Tables 6.6 and 6.7 with those in Table 6.5, it can be seen that recognition rates are lowest for the "realistic" HMM2 features. Classification errors made by HMM2 in the first pass and classification errors made by the second-pass HMM using error-prone HMM2 features accumulate.

As already discussed before, the use of HMM2 features as such might be questioned because the recognition rates of HMM2 features used in a second recognition pass (Table 6.5) are worse than those of an HMM2 system directly applied as a decoder (Table 6.6). However, there are several arguments in favour of using an HMM2 system for feature extraction. Firstly, HMM2 features may be useful for the analysis of speech signals. Secondly, in spite of their obvious disadvantages, their classification performance is competitive with that of RF, a more common example of automatically extracted formant related features, and in the case of the gender-dependent tests even with that of HLF. Thirdly, although HMM2 features, RF and even HLF achieve worse results than state-of-the-art features such as MFCCs, formant related features can be supposed to represent complementary information, and can therefore be used not instead of but in addition to these state-of-the-art features, especially in difficult conditions (as has already been shown in Section 5.4.3)[8].

b) Visual comparison

In the following, we will give a visual comparison of hand-labeled formant tracks and HMM2 feature tracks obtained from these models. Figure 6.6 shows HLF tracks in the left and HMM2 feature tracks (of the same speech segment) in the right part of each sub-figure, both projected onto the corresponding spectrograms of FF2 features. While the HLF tracks are denoted F1 (for the formant track in the lowest frequencies), F2 and F3, the HMM2 feature tracks will be called T1, T2 and T3 respectively. Generally, it has to be stated that the correspondence between HMM2 feature tracks and hand-labeled formants is marginal. However, in many cases it can be seen that the HMM2 model was able to separate high delta-energy regions from low ones. As the underlying features are frequency derivatives (i.e., delta-energies, representing changes of energy along the frequency axis), the transitions from one frequency HMM state to the next may correspond to spectral peaks or valleys of the original spectral features, as was argued in Section 6.2.1.

Let us consider for example a pronunciation of the vowel /er/ as shown in Figure 6.6(a)[9]. Starting from the lowest frequencies, there is an alternation of increasing and decreasing energies. In the case of HMM2 (right panel), the frequency HMM states 1 and 3 seem to model positive changes in energy, while states 2 and 4 model negative changes. T1 (i.e., the transition between state 1 and state 2) would thus correspond to a spectral maximum, which, as argued further above and visualized in Figures 6.4, could be related to a formant. Indeed, a visual comparison between T1 and F1 (in the HLF figure, left panel) shows that their positions are rather close. While T2 corresponds to a spectral valley, T3 is again close to the position of an HLF track (in this case, F2). Even a small upward tendency of F2 towards the

[8]However, contrary to the results reported in Section 5.4.3, an improved robustness for speech degraded by additive noise was not observed in our experiments with the AEV database (de Wet et al., 2002).

[9]This is in fact the same example as in Figures 6.2(a), 6.4, 6.5.

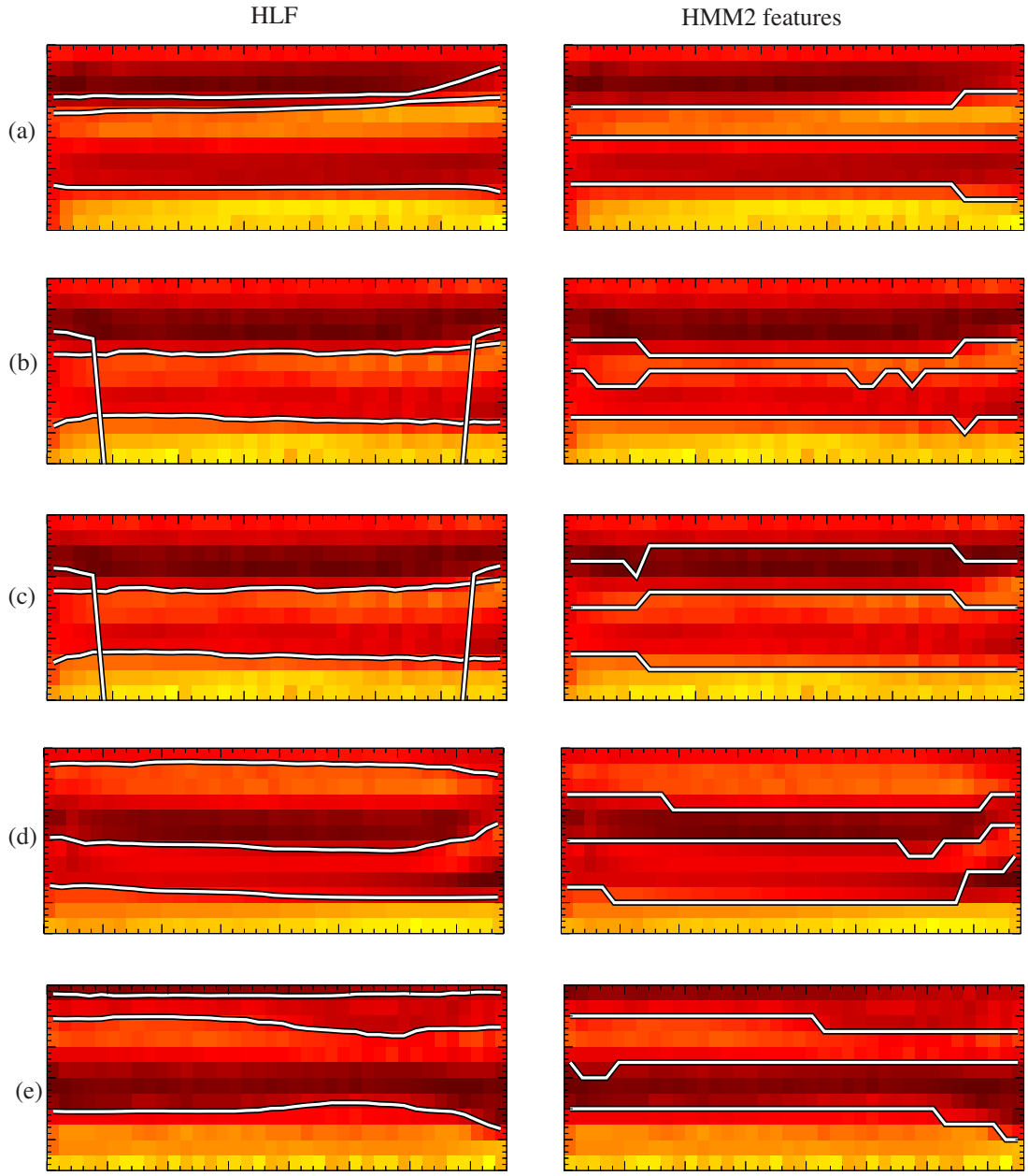


Figure 6.6: Comparison of hand-labeled formant tracks (left column) and HMM2 feature tracks (right column), overlaid onto a spectrogram of FF2 features. In (a), (b), and (c), examples of phoneme /er/ are shown, (d) figures an example of /oa/ and (e) of /ae/. The HMM2 feature tracks of (a) and (b) were obtained using gender-dependent models, those of (c), (d), and (e) using gender-independent models. (b) and (c) are showing the same example for a direct comparison between HMM2 feature tracks obtained from gender-dependent and gender-independent models.

end of the speech segment is reflected by T3 (although T3 is much more crude than the smooth formant track). F3 is not represented by the HMM2 feature tracks. In fact, F2 and F3 are quite close, and the low frequency resolution of the FF2 features does not permit to distinguish these two formants. Figure 6.6(b) shows a different pronunciation of phoneme /er/, and it can be seen in the HLF tracks that a

merger between F2 and F3 occurred. For the HMM2 tracks, T1 and T3 again reflect approximately formant positions. As compared to example (a), the frequencies of F2 are slightly lower, and also T3 is situated one feature below the T3 track in Figure (a) for most of its duration. The T2 track is quite irregular: oscillations occur around a certain frequency band, which is sometimes assigned to state 2 or state 3 respectively. This effect might for example occur when the "right" track would be situated just between two discrete frequency positions.

While the HMM2 features tracks in Figure 6.6(a) and (b) originated from gender-dependent models, those in (c), (d) and (e) were obtained using gender-independent models. Figures 6.6(b) and (c) show the same pronunciation, but the HMM2 feature tracks in (c) do not correspond to what would have been expected in terms of separation between high and low energies, and was obtained in (b). This discrepancy might be explained by the larger variation of formant positions in the gender-independent data, and therefore a higher potential of confusion.

Figure 6.6(d) shows an example pronunciation of phoneme /oa/. While T2 is quite similar to F2, F1 and F3 are not so well represented. Intuitively, F3 can not be well derived from the FF2 data used, and the T3 track (which seems to follow a spectral valley) seems indeed to make more sense (given these features). While the first part of T1 seems to approximately correspond to F1, there is a transition to much higher frequencies at the end of the speech segment. This may be explained with a transition from one temporal HMM state to the next, which sometimes accounts for rather unexpected transitions (as also seen, e.g., in Figure 6.6(c)). In Figure 6.6(e), yet another example (of phoneme /ae/) is shown, where T1 is close to F1 and T3 to F2 respectively.

Summarizing the HMM2 feature tracks in the figure, it can be stated that often one or two tracks correspond to a certain degree to hand-labeled formants. Another HMM2 feature track frequently follows a spectral valley. However, the accuracy of the HMM2 feature tracks is severely limited by the low frequency resolution of the FF2 features employed.

6.4.4 Summary of Results and Discussion

Figure 6.7 visualizes the most important results obtained on the AEV database. The left cluster shows the results for the gender-independent (GI), and the right cluster for the gender-dependent (GD) tests. Generally, it can be stated that the performance of the GD tests is higher than that for GI. However, the differences between GI and GD results are comparatively small for the case of MFCC-13, suggesting that these features are well-suited for modeling speech rather than speaker dependent characteristics of the signal. This is however not the case if only the first 3 MFCC features are used, where the differences between GI and GD results are rather large. Large differences for GI vs. GD performance could also be observed for both automatically extracted formant related features.

For each cluster, 95% confidence intervals are given with respect to the HLF results, as the aim of these experiments was a comparison based on the performance of hand-labeled formants. It can be seen that MFCC-13 perform significantly better than HLF, for both gender-independent and gender-dependent models. For the gender-independent case, HLF achieve significantly higher classification rates than both automatically extracted features. On the other hand, for the gender-dependent case, RF as well as HMM2 features are comparable to HLF.

The generally poorer performance (especially of formant related features) in the GI as compared to the GD experiments can be explained by a greater overlap between features belonging to different phonemes, and thus a higher confusability. In addition to that, a priori knowledge about the gender may be used during feature extraction. In fact, for the case of HLF it is likely that the human labeler made (con-

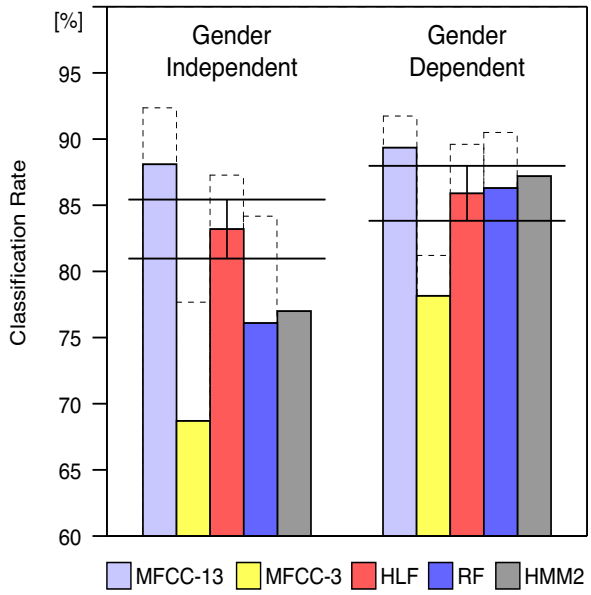


Figure 6.7: Summary of important results. The left cluster shows average classification rates for the gender-independent tests, the right cluster for the gender-dependent ones. The bars in each cluster correspond to the following features (from left to right): MFCC-13, MFCC-3, HLF, RF, and HMM2 features. Moreover, where appropriate, results using the same features with additional first order temporal derivatives are indicated with broken lines. The errorbars shown for each cluster are based on the HLF results and indicate the 95% confidence interval.

sciously or not) use of such a priori knowledge (and even of knowledge concerning the phoneme identity, possibly resolving ambiguities in candidate formant tracks). Therefore, the HLF features might be extracted in a gender-dependent way, which would positively influence the classification performance also of the GI models. In contrast, for the GI tests on the RF and HMM2 features, such a priori knowledge was supposed to be unavailable, resulting in a gender-independent way of extracting these features. Therefore, for each of the automatically extracted formant related features, there are two different sets of features, i.e., a gender-dependent and a gender-independent set. Consequently, the features used for the GI tests are likely to be sub-optimal as compared to those used in the GD tests. This adds up with a greater overlap between different phonemes, as discussed above. These two reasons might explain the rather large differences for GI and GD experiments for the case of automatically extracted formant related features, as compared to the HLF.

Comparing RF and HMM2 features, it can be stated that the differences in the classification rates are minor for both the GI and the GD tests. However, this is only true for the case where no temporal derivatives are appended (de Wet, 2002). When using additional first-order temporal derivatives, RF performance increases significantly. In contrast, due to the very crude nature of HMM2 features (as demonstrated in Figure 6.6), temporal derivatives are meaningless for this case. However, it should be noted that this is a consequence of the present choice of features and HMM2 system topology, and may not necessarily generalize to other HMM2 feature extractor implementations.

6.5 Conclusion

The principal goal of this chapter was to investigate the vowel classification performance of HMM2 features as compared to state-of-the-art features, hand-labeled formants, and other automatically extracted formant related features. It can be stated that (higher-dimensional) MFCCs outperform all kinds of formant related features. Considering that the study presented in this chapter was limited to a vowel classification task (due to the limitations imposed by the database), it might be expected that MFCCs even more significantly outperform formant-like features for the case of consonants. Therefore, the results are rather theoretical, and can be seen as an upper limit of the performance of formant related features under ideal, appropriate conditions.

Given these limitations, very encouraging results were observed. In particular, comparable classification rates were obtained in the case of gender-dependent models for hand-labeled formants, robust formants and HMM2 features. However, a visual comparison between these features demonstrates the limits of the current implementation of the HMM2 system. In fact, the HMM2 feature tracks are very crude, and typically at least one of the three tracks follows a spectral valley rather than a spectral peak. For these reasons, the correspondence to real formant tracks is rather limited. The competitive classification results obtained from HMM2 features suggest however that the consistency of the features is possibly more important than an exact correspondence to formant values.

It is however still an open research issue if even better results could be obtained if HMM2 features were more formant-like. In fact, the HMM2 system could be changed in order to obtain HMM2 features which closely resemble formant tracks. Firstly, a more appropriate signal representation should be chosen, featuring a higher frequency resolution. It is a clear priori that using only 12 FF features will not allow to extract real formant features. Instead, a higher number of narrower sub-bands should be used. Possibly, a more sophisticated signal processing method is needed than simple filterbank differences. However, contrary to the benefit of a better frequency resolution, the higher feature dimension might also cause problems, such as a much higher complexity during HMM2 training and recognition / features extraction.

A better frequency resolution is likely to result in a higher number of spectral peaks and valleys. For instance, two formants close in frequency (i.e., which are represented by a single spectral peak in the present features) could be resolved. This necessitates an adapted the HMM2 system topology. The number of frequency HMM states should be increased, resulting in a higher number of frequency segmentations. Of these, the ones corresponding to spectral valleys could be discarded, only using the ones corresponding to spectral peaks (which are obviously more formant-like) as features. Additionally, a more sophisticated, possibly even phoneme-dependent frequency HMM topology could be employed, in combination with an adequate post-processing of the resulting frequency segmentations in order to extract the relevant formant-like feature tracks.

Chapter 7

Conclusions and Outlook

7.1 General Summary

In this thesis, the modeling of temporal and frequency correlation of speech signals by HMMs, with application to robust automatic speech recognition, was investigated. Particular attention was devoted to the HMM2 approach, where the HMM paradigm (on which the temporal modeling of speech signals is typically based) was extended towards the frequency dimension of speech. The most important results achieved in this thesis are summarized in the following.

- Information from relatively long time scales, appended to conventional feature vectors, can be used to improve ASR robustness (Weber, 2000). This can be seen as a shift of the modeling of temporal correlation from the HMM part towards the GMM part of the GM-HMM. As these long-term features generally do not provide any additional information as compared to a sequence of conventional, temporal feature vectors, it can be supposed that the performance improvements observed are related to a better modeling of this temporal correlation in GMMs.
- On the other hand, the inverse approach consists of splitting up each temporal feature vector into a sequence of sub-vectors. Instead of modeling the entire original vector by a GMM, this sequence of sub-vectors can then be modeled by a (secondary) HMM, where the emission distributions are again represented by (low-dimensional) GMMs. This is the essential idea of the HMM2 approach (Weber, Bengio, and Bourlard, 2000). First experiments related to HMM2 were done using wavelet features and an adapted, tree-like secondary HMM (Keller, Ben-Yacoub, and Mokbel, 1999). Other variants include ergodic and trellis-like secondary HMMs (Weber, Bengio, and Bourlard, 2000). However, further research focused on a particular HMM2 implementation, using filterbank-based spectral features and a bottom-up looped secondary HMM topology (Weber et al., 2002).
- Potential advantages of HMM2 include the modeling of correlation through the secondary HMM topology and a better and more flexible modeling and parameter sharing. Moreover, the non-linear state-dependent spectral warping performed by the secondary HMM could be useful for an (implicit or explicit) vocal tract normalization (Ikbal, Weber, and Bourlard, 2002). Also, HMM2 can be seen as a more flexible extension to multi-band processing (Bourlard, Bengio, and Weber, 2001; Bourlard, Bengio, and Weber, 2002).
- The EM algorithm, conventionally used to train HMMs, can be adapted to the case of HMM2 (Bengio, Bourlard, and Weber, 2000). An HMM2 system can be implemented either using this

adapted EM algorithm (Ikbal et al., 2001), or by using conventional EM training on a large, "unfolded" HMM2 with additional synchronization constraints (Weber, Bengio, and Bourlard, 2001b).

- From a practical point of view, some attention has to be devoted to the realization of an HMM2 system. In particular, an additional frequency coefficient, appended to each feature sub-vector, was shown to be useful (Weber, Bengio, and Bourlard, 2001c). Moreover, as for conventional HMMs, the choice of hyper-parameters is important. Care has to be taken in order to set appropriate minimum Gaussian variances, and to choose a suitable initialization procedure (Weber et al., 2002).
- When using HMM2 for the recognition of clean speech, performance drops were observed as compared to standard HMMs. This suggests that the correlation between the coefficients of a feature vector (e.g., corresponding to correlation in frequency) can be more efficiently modeled by GMMs than by HMMs. This experimental result also confirms the results of theoretical investigations made on this subject (Weber, Bengio, and Bourlard, 2001b).
- In the case of noisy speech, the HMM2 approach was shown to outperform conventional GM-HMMs when using the same spectral features for both systems (however, it should be noted that this comparison is not completely fair, as HMMs using MFCCs and noise reduction techniques yield better results). This suggests that the modeling of frequency correlation in GMMs (i.e., the conventional way) is suitable for matched training and testing conditions, but might be sub-optimal where there is a mismatch between training and test data. In the case of mismatch, a more flexible approach to modeling, such as that provided by HMM2, seems more appropriate (Weber, Bengio, and Bourlard, 2002).
- HMM2 was also shown to model pertinent structural information of the speech signal (Weber, Bengio, and Bourlard, 2000; Weber, Bengio, and Bourlard, 2001c). This information can be extracted and converted to features that are useful for ASR. While the recognition performance obtained using these "HMM2 features" is not competitive with systems using state-of-the-art MFCCs, the combination of these different features in a multi-stream approach has led to improvements in ASR robustness (Weber, Bengio, and Bourlard, 2001a; Weber, Bengio, and Bourlard, 2002). This result (again) indicates that the HMM2 model may provide the flexibility necessary for dealing with mismatched conditions.
- A further comparison between HMM2 features and other formant-related features (and additionally with MFCCs) in terms of their vowel classification performance was motivated by the assumption that the structural information modeled by HMM2 could be related to formants. While all formant-related features yielded results inferior to MFCCs, the differences between HMM2 features, hand-labeled formants and other formant-related features were not significant. This suggests that, although the HMM2 features tested here do not generally correspond to true formants, the information content of these different features in terms of their capacity to discriminate between different speech sounds is comparable (Weber et al., 2002; de Wet et al., 2003).
- Although the focus of this thesis was not on multi-stream processing, it can be confirmed from several experiments that different kinds of additional features, if combined with conventional state-of-the-art features at either the feature or the local likelihood level, can enhance ASR robustness (Keller, Ben-Yacoub, and Mokbel, 1999; Weber, 2000; Weber, Bengio, and Bourlard, 2001a; etc.).

7.2 Future Directions Towards a More Flexible Modeling of Speech

The HMM2 approach, as a generalization of the conventional GM-HMM systems, offers a powerful framework for the modeling of the variability inherent in the (possibly degraded) speech signal, which has not yet been fully investigated. Consequently, there is still much scope for improvement, and for discovery of promising new fields of application. Several possible directions for future research are outlined below.

First and foremost, future investigations related to the work presented in this thesis in general and to the HMM2 approach in particular should aim at finding a better trade-off between the modeling of temporal and frequency correlation in HMMs vs. GMMs. The results presented in this thesis indicate that, while the modeling of sequences of coefficients by GMMs is suitable in many cases, HMMs may be especially advantageous for the case where variability in the data is high and unpredictable (e.g., where training can not be done on all possible conditions of data variability). The HMM2 approach offers a framework for investigating these issues, allowing a shift of the modeling by GMMs further towards HMMs.

A first research issue is related to the choice of features in general. Staying in the spectral domain, it might be advantageous to consider a higher degree of detail in the features, pointing towards larger feature vectors offering a finer frequency resolution. On the other hand, it could be considered to include more temporal information into the feature vector. While research results (presented in this thesis, but also by other researchers) indicate advantages of using multi-resolution features, the optimal trade-off between the temporal and frequency resolution is not yet clear. This question might be even more crucial for the particular case of HMM2. Here, it should further be determined which information needs to be represented by a single feature vector, i.e., at which level the (original, temporal) feature vector should be split up into smaller sub-vectors, which are then modeled by a GMM. Selection criteria might be the degree of correlation, or the mutual information, between different coefficients.

Secondly, and closely related to the choice of features and the partition of coefficients into distinct vectors, is the search for an adapted HMM2 topology. This includes more complex (and possibly phoneme-dependent) structures within the secondary HMM, but also a more sophisticated modeling of correlation in both the temporal and frequency dimensions could be considered. Moreover, similar to hybrid HMM/ANN systems, ANNs could be advantageous for modeling the local emission probabilities associated with the HMM2 states.

Given such optimizations to the HMM2 system, it could also be used for a more sophisticated (and possibly more formant-related) modeling of the speech signal, improving also its application as a feature extractor. An additional research issue especially related to this application would be to determine what kind of speech unit should be modeled (e.g., considering broader classes than phonemes or subphone units, thereby diminishing the problem of inaccuracies of HMM2 features due to misclassifications performed by the HMM2 systems).

The HMM2 approach also opens up new perspectives for multi-band processing, e.g., allowing for a dynamic and data-dependent definition of sub-bands, an issue which has been only briefly touched upon in this thesis. Another related research issue might be concerned with the relationship between HMM2 and missing data processing, possibly leading to new modeling variants which combine the advantages of these two approaches. Furthermore, the HMM2 framework could provide new methods for speaker adaptation (e.g., based on a non-linear vocal tract normalization, or using a flexible and adaptive, speaker (and phoneme) dependent variants of the Mel scale (as a function of the frequency warping per-

formed by HMM2). These issues are in fact currently under investigation (Ikbal, Weber, and Bourlard, 2002). Finally, the HMM2 approach could possibly be adapted to speaker recognition or verification.

7.3 Final Thoughts

When I started working on this thesis, I had the choice to either continue on well-beaten tracks of speech recognition research, or to pursue more innovative and novel ideas. The former option appeared to have the advantage of offering higher chances of fast success, while the latter seemed to be more adventurous and risky, but infinitely more interesting (Mokbel, 1999). Following the spirit of (Bourlard, Hermansky, and Morgan, 1996), I chose the second option, and I indeed succeeded in improving speech recognition error rates in the first place. While a lot of progress has been made since these early days, the work accomplished in this thesis can only be seen as a first step, in which a novel approach was only touched upon and as yet remains far from being exhaustively explored. Nor is the list of possible future research directions intended to be complete. In this sense, I firmly hope that this thesis might give new inspirations to speech recognition (and possibly other fields of) research, whether or not they are based on the exact methods presented here.

"We cannot reach new horizons if we fear to leave the shore."

Appendix A

Results for HMM2 Decoder

In the following, more results on the application of HMM2 as a decoder are given. This completes the results given in Section 4.2.2. As features, second order frequency filtered filterbanks were used. The tables below report word error rates (WER) obtained on the Numbers95 database with different additive noises on different signal-to-noise ratios (SNR). Results obtained on HMM2 are compared to those obtained using conventional HMMs (baseline system).

HMM Baseline	HMM2
6.4	13.2

Table A.1: Comparison of HMM2 decoder performance: WER on clean speech.

SNR	HMM Baseline	HMM2
18	8.6	15.3
12	14.2	20.7
6	31.7	35.7
0	64.9	62.4

Table A.2: Comparison of HMM2 decoder performance: WER on factory noise.

SNR	HMM Baseline	HMM2
18	7.9	14.4
12	12.4	16.9
6	26.8	26.1
0	56.1	45.8

Table A.3: Comparison of HMM2 decoder performance: WER on lynx noise.

SNR	HMM Baseline	HMM2
18	9.3	14.8
12	16.7	19.4
6	36.8	32.3
0	66.2	56.1

Table A.4: Comparison of HMM2 decoder performance: WER on car noise.

Appendix B

Results for HMM2 Feature Extractor

In the following, more result on the application of HMM2 as a feature extractor are given. This completes the results given in Section 5.4.3. The tables below report word error rates (WER) obtained on the Numbers95 database with different additive noises on different signal-to-noise ratios (SNR). The baseline system uses MFCCs (including cepstral mean subtraction and spectral subtraction) as features. The HMM2 features were obtained from the OM system. Moreover, results obtained when combining MFCCs with HMM2 features are shown.

MFCC-SS (baseline)	HMM2 features (OM)	MFCC-SS + HMM2 features
5.7	43.2	5.6

Table B.1: Comparison of HMM2 feature performance: WER on clean speech.

SNR	MFCC-SS (baseline)	HMM2 features (OM)	MFCC-SS + HMM2 features
18	7.4	42.3	7.3
12	11.9	49.8	11.4
6	23.0	62.2	21.4
0	48.6	76.4	46.6

Table B.2: Comparison of HMM2 feature performance: WER on factory noise.

SNR	MFCC-SS (baseline)	HMM2 features (OM)	MFCC-SS + HMM2 features
18	6.2	41.0	6.0
12	7.4	43.5	7.4
6	12.3	50.5	12.1
0	24.2	62.6	23.6

Table B.3: Comparison of HMM2 feature performance: WER on lynx noise.

SNR	MFCC-SS (baseline)	HMM2 features (OM)	MFCC-SS + HMM2 features
18	6.6	41.7	6.2
12	8.6	45.8	8.8
6	14.7	56.7	14.8
0	33.5	70.2	33.5

Table B.4: Comparison of HMM2 feature performance: WER on car noise.

Appendix C

Results on the American English Vowels Database

In the following, a summary of preliminary results obtained on the American English Vowels database is given. Different design, initialization and training/testing options, such as described in Section 6.3.4, have been tested. While the most important results are outlined in Section 6.4.3, some more details will be given below. However, first the abbreviations as used in the tables are summarized.

Abbreviations

OM	one model comprising one temporal HMM state, trained on all data
PDM-1	phoneme-dependent models, each comprising one temporal HMM state
PDM-3	phoneme-dependent models, each comprising three temporal HMM states
LHLH	initialization assuming alternating high and low energy bands, starting with low energy in the lowest frequencies
HLHL	initialization assuming alternating high and low energy bands, starting with high energy in the lowest frequencies
FMT	initialization considering formant frequencies for the frequency segmentation
Lin	initialization assuming a linear segmentation in frequency
FC	feature combination between frequency coefficient and other features
LC	likelihood combination between frequency coefficient and other features
FA	forced alignment for HMM2 feature extraction
RC	recognition for HMM2 feature extraction

a) OM and PDM systems

Model topology	Frequency Coefficient	Initialization	Training/ Testing	All	Male	Female
OM	FC	LHLH	FA/RC	33.0	44.8	40.2
PDM-1				65.2	82.6	81.9
PDM-3				77.0	82.4	83.7

Table C.1 Comparison (classification rates) of OM and PDM systems.

b) Frequency Coefficient

Model topology	**Frequency Coefficient**	Initialization	Training/ Testing	All	Male	Female
OM	**FC**	LHLH	---	33.0	44.8	40.2
	LC			61.9	76.5	57.2
PDM-1	**FC**		FA/RC	65.2	82.6	81.9
	LC			68.6	79.4	78.5
PDM-3	**FC**			77.0	82.4	83.7
	LC			75.8	79.4	82.4

Table C.2 Comparison (classification rates) of using the frequency coefficient in feature combination (FC) vs. likelihood combination (LC). For the case of LC, different stream weights have been tested and only the best results are reported here.

c) Initialization

Model topology	Frequency Coefficient	**Initialization**	Training/ Testing	All	Male	Female
PDM-3	FC	**Lin**	FA/RC	70.0	77.0	86.3
		LHLH		77.0	82.4	83.7
		HLHL		71.7	74.8	86.3
		FMT		71.2	84.4	90.0

Table C.3 Comparison (classification rates) of different initialization methods.

d) Training and Testing

Model topology	Frequency Coefficient	Initialization	**Training/ Testing**	All	Male	Female
PDM-3	FC	LHLH	**RC/RC**	74.4	82.8	83.0
			FA/RC	77.0	82.4	83.7
			(FA/FA)	81.6	89.1	88.9
		FMT	**RC/RC**	71.2	84.4	89.4
			FA/RC	71.2	84.4	90.0
			(FA/FA)	77.9	94.1	93.0

Table C.4 Comparison (classification rates) of different options for training/testing.

Notations

A	auxiliary function
a	transition probability, where a_{ij} is the transition probability in the primary HMM (to go from state i to state j), and a_i^{lm} is the transition probability in secondary HMM associated with primary HMM state i (to go from secondary state l to secondary state m)
b	frequency band
c	mixture weight, where c_{ilg} is the weight of the g-th mixture component associated with primary HMM state i and secondary state l
d	dimension (of a feature vector)
F	length of sequence of (frequency) sub-vectors
f	frequency, where f_l and f_h denote low and high cut-off frequencies respectively, and f_{max} is the maximum frequency
FI	frequency index
G	number of Gaussian mixtures
g	g-th mixture component
i, j	designate a primary HMM state
k	k-th training iteration
l, m	designate a secondary HMM state
N	number of primary states
N_i	number of secondary states in primary state i
P	probability
p	probability density function
Q	a path in the primary HMM
Q^*	best path in the primary HMM
Q_t	a path in the secondary HMM associated with temporal HMM state visited at time t
Q^+	a path in the HMM2 (primary and secondary HMM)
q_t	primary HMM state at time step t
q_t^f	secondary HMM state at time step t and frequency f
s	stream

T	length of acoustic feature vector sequence
t	time, where t_s and t_e denote the start and stop time of a speech unit respectively
TI	time index
w	
x_t	observed feature vector at time step t
$x_{1,T}$	observed feature vector sequence from time step 1 to T
x_t^f	observed feature component at time t and frequency f
$x_t^{1,F}$	equivalent to x_t
$\overline{x_{t(b)}}$	mean of components of frequency band b at time t
Z, Z'	indicator variables
θ	parameter set
μ	mean, where μ_{ilg} is the mean of g-th Gaussian mixture of the i-th temporal and the l-th frequency HMM state
σ^2	variance, where σ_{ilg}^2 is the variance of g-th Gaussian mixture of the i-th temporal and the l-th frequency HMM state

Abbreviations

A	Acceleration (or delta-delta) feature coefficient
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
C	Feature Coefficient
CI	Confidence Interval
CMS	Cepstral Mean Subtraction
D	Delta feature coefficient
EM	Expectation Maximization algorithm
FF	Frequency Filtered filterbanks
FF2	second order frequency Filtered Filterbanks
FI	Frequency Index
F1, F2, F3	first three Formants
GMM	Gaussian Mixture Model
GM-HMM	HMM employing a GMM for phoneme emission probability (likelihood) estimation
HMM	Hidden Markov Model
HMM/ANN	HMM employing an ANN for phoneme emission probability estimation
HMM2	HMM employing an HMM for phoneme emission probability estimation
LDA	Linear Discriminant Analysis
MFCC	Mel Frequency Cepstral Coefficient
OM	One Model variant of HMM2
PCA	Principal Component Analysis
PDF	Probability Density Function
PDM	Phoneme-Dependent Model
RF	Robust Formants
SLA	Split Levinson Algorithm
SNR	Signal-to-Noise Ratio
SS	Spectral Subtraction
TI	Time Index
WER	Word Error Rate
WHMT	Wavelet-domain Hidden Markov Tree

Bibliography

Acero, A. (1999). Formant analysis and synthesis using hidden Markov models. In *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, volume 1, pages 1047-1050.

Antonini, M., Barlaud, M., Mathieu, P., and Daubechies, I. (1992). Image Coding Using Wavelet Transform. *IEEE Transactions on Image Processing*, volume 1, number 2, pages 205-220.

Atal, B. S. and Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, volume 50, pages 637-655.

Bahl, L. R., Brown, P. F., de Souza, P. V., and Mercer, R. L. (1986). Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 49-52.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 37:1554-1563.

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical statistics*, 41:164-171.

Bengio, S., Bourlard, H. and Weber, K. (2000). An EM Algorithm for HMMs with Emission Distributions Represented by HMMs. Technical Report IDIAP-RR 00-11.

Bilmes, J.A. (1999a). *Natural Statistical Models for Automatic Speech Recognition*, Ph.D. Dissertation, Dept. of EECS, University of California, Berkeley.

Bilmes, J.A. (1999b). Buried Markov Models for Speech Recognition. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 713-716.

Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

Bourlard, H., Bengio, S., and Weber, K. (2001). New Approaches Towards Robust and Adaptive Speech Recognition. In *Advances in Neural Information Processing Systems 13,* MIT Press.

Bourlard H., and Bengio, S. (2002). Hidden Markov Models and other Finite State Automata for Sequence Processing. In *The Handbook of Brain Theory and Neural Networks: The Second Edition*, 2002.

Bourlard, H., Bengio, S., and Weber, K. (2003). Towards Robust and Adaptive Speech Recognition Models. To appear in Ostendorf, M., Khudanpur, S., and Rosenfeld, R., editors, *Mathematical Foundations of Speech Processing and Recognition*, Institute for Mathematics and its Applications (IMA) Series. Springer-Verlag.

Bourlard, H. and Dupont, S. (1996). A New ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 422-425.

Bourlard, H. and Morgan, N. (1994). *Connectionist Speech Recognition. A Hybrid Approach*. Kluwer Academic Publishers, Boston MA.

Bourlard, H., Hermansky, H., and Morgan, N. (1996). Towards Increasing Speech Recognition Error Rates. Speech Communication, volume 18, pages 205-231.

Bronstein, I.N. and Semendjajew, K.A. (1989), *Taschenbuch der Mathematik*. Verlag Nauka, Moskau, BSB B. G. Teubner Verlagsgesellschaft, Leipzig.

Choi, H. and Baraniuk, R.G. (1999). Image Segmentation using Wavelet-domain Classification. In *Proc. SPIE Technical Conference on Mathematical Modeling, Bayesian Estimation, and Inverse Problems*, pages 306-320.

Cole, R. A., Noel, M., Lander, T., and Durham, T. (1995). New Telephone Speech Corpora at CSLU. In *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, volume I, pages 821-824.

Cooke, M.P., Green, P.D., Josifovski, L., and Vizinho, A. (2001). Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data. Speech Communication, volume 34, number 3.

Crouse, M. S., Nowak, R. D., and Baraniuk, R. G. (1998). Wavelet-Based Statistical Signal Processing Using Hidden Markov Models. *IEEE Transactions on Signal Processing*, volume 46, number 4, pages 886-902.

Daubechies, Ingrid (1992). Ten Lectures on Wavelets. *CBMS-NSF Regional Conference Series in Applied Mathematics*, volume 61, Society for Industrial and Applied Mathematics (SIAM), Philadelphia.

Davis, S.B. and Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, volume 28, pages 357-366.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum-likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society* B, volume 39, pages 1-38.

DeVore, R., Jawerth, B., and Lucier, B. (1992). Image compression through wavelet transform coding. *IEEE Transactions on Information Theory*, volume 38, number 2, pages 719-746.

Duda, R. O., and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. New York, NY: Wiley.

Eickeler, S., Müller, S. and Rigoll, G. (1999). High Performance Face Recognition Using Pseudo 2D-Hidden Markov Models. In *Proc. European Control Conference (ECC)*.

Ellis, D. P. W. (2000). Stream combination before and/or after the acoustic model. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 1635-1638.

Farooq, O. and Datta, S. (2001) Mel Filter-Like Admissible Wavelet Packet Structure for Speech Recognition. *IEEE Signal Processing Letters*, volume 8, number 7, pages 196-198.

Fukunaga, K. (1972). *Introduction to Statistical Pattern Recognition*. Academic Press, New York.

Gales M.J.F. and Young S.J. (1993). Segmental HMMs for Speech Recognition. In *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1579-1582.

Gales M.J.F. and Young S.J. (1995). Robust Speech Recognition in Additive and Convolutional Noise Using Parallel Model Combination. *Computer Speech and Language*, 9:289-307.

Gales M.J.F. and Young S.J. (1996). Robust Continuous Speech Recognition using Parallel Model Combination. *IEEE Transactions on Speech and Audio Processing*, volume 4, number 5, pages 352-359.

Garner, P. and Holmes, W. (1998). On the Robust Incorporation of Formant Features into Hidden Markov Models for Automatic Speech Recognition. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 1-4.

Ghahramani Z. and Jordan, M.I. (1997). Factorial hidden Markov models. *Machine Learning*, 29:245-273, 1997.

Glotin, H. (2000). *Elaboration et compairison de systèmes adaptatifs multi-flux de reconnaissance robuste de la parole: incorporation des indices d'harmonicité et de localisation*. Ph.D. thesis, Institut National Polytechniques de Grenoble, France.

Gold, B. and Morgan, N. (2000). *Speech and Audio Signal Processing*. John Wiley & Sons, Toronto.

Gravier, G., Sigelle, M., and Chollet, G. (2000). A Markov Random Field Based Multi-band Model. *In Proc. International Conference on Acoustics, Speech and Signal Processing*.

Haeb-Umbach R. and Ney, H. (1992). Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition. *In Proc. International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 13-16.

Hagen, A. (2001). *Robust Speech Recognition Based on Multi-Stream Processing*. Ph.D. thesis, Departement d'informatique, Ecole Polytechnique Federale de Lausanne, Switzerland.

Hasegawa-Johnson, M. (1996). *Formant and Burst Spectral Measurements with Quantitative Error Models for Speech Sound Classification*. Ph.D. thesis, MIT, Cambridge, MA, August 1996.

Hermansky, H., Morgan, N., Bayya, A., and Kohn, P. (1991). RASTA-PLP Speech Analysis. Technical Report (TR-91-069), International Computer Science Institute, Berkeley, CA.

Hermansky, H. and Morgan, N. (1994). Rasta Processing of Speech. *IEEE Transactions on Speech and Audio Processing*, special issue on Robust Speech Recognition, volume 2, number 4, pages 578-589.

Hermansky, H. and Sharma, S. (1999). Temporal Patterns (TRAPS) in ASR of Noisy Speech. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 289-292.

Hermansky, H., Ellis, D. and Sharma, S. (2000). Tandem Connectionist Feature Extraction for Conventional HMM Systems. In *Proc. International Conference on Acoustics, Speech and Signal Processing*.

Hillenbrand, J., Getty, L.A., Clark, M.J., and Wheeler, K. (1995). Acoustic Characteristics of American English Vowels. *Journal of the Acoustical Society of America*, volume 97, number 5, pages 3099-3111.

Holmes, J., Holmes, W., and Garner, P. (1997). Using Formant Frequencies in Speech Recognition. In *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, volume 4, pages 2083-2086.

Holmes, W. (2000). Segmental HMMs: Modelling Dynamics and Underlying Structure for Automatic Speech Recognition. *IMA Workshop on Mathematical Foundations of Speech Processing and Recognition,* workshop material, http://www.ima.umn.edu/multimedia/fall/m1.html.

Huang, X., Acero, A., and Hon, H. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall, Upper Saddle River, New Jersey.

Hunt, M. and Lefebvre, C. (1989). A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 262-265.

Hunt, M.J. and Richardson, S.M. (1990). Use of Linear Discriminant Analysis in a Speech Recognizer. *Speech Tech Worldwide*, pages 87-93.

Hunt, M.J. (1987). Delayed Decisions in Speech Recognition - the Case for Formants. *Pattern Recognition Letters*, 6:121-137.

Ikbal, S., Bourlard, H., Bengio, S. and Weber, K. (2001). IDIAP HMM/HMM2 System: Theoretical Basis and Software Specifications. Technical Report IDIAP-RR 01-27.

Ikbal, S., Weber, K., and Bourlard, H. (2002). Speaker Normalization using HMM2. In *Proceedings of the International IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, Martigny Switzerland, 2002.

Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. Language, Speech and Communication Series. MIT Press, Cambridge, MA.

Junqua, J. and Haton, J. (1996). *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. Kluwer Academic Publisher, Boston.

Kadambe S. and Srinivasan, P. (1994). Applications of Adaptive Wavelets for Speech. *Optical Engineering*, 33(7):2204-2211.

Keller (now Weber), K., Ben-Yacoub, S., and Mokbel, C. (1999). Combining Wavelet-domain Hidden Markov Trees with Hidden Markov Models. Technical Report IDIAP-RR 99-14.

Kermorvant, C. (1999). A Comparison of Noise Reduction Techniques for Robust Speech Recognition. Technical Report IDIAP-RR- 99-10.

Kermorvant, C. and Morris, A. (1999). A comparison of Two Strategies for ASR in Additive Noise: Missing Data and Spectral Subtraction. In *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 2891-2844.

Kopec, G. (1986). Formant Tracking using Hidden Markov Models and Vector Quantization. *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 34, pages 709-729.

Kryze, D., Rigazio, L., Applebaum, T., and Junqua, J. (1999). A New Noise-Robust Subband Front-End And Its Comparison To PLP. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Kuo, S. and Agazzi, O. (1993). Machine Vision for Keyword Spotting Using Pseudo 2D Hidden Markov Models. *In Proc. International Conference on Acoustics, Speech and Signal Processing*, volume V, pages 81-84.

Laprie, Y., and Berger, M. (1994). A New Paradigm for Reliable Automatic Formant Tracking. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 201-204.

Lee, L. and Rose, R. (1998). A Frequency Warping Approach to Speaker Normalization. *IEEE Transactions on Speech and Audio Processing*, volume 6, number 1, pages 49-59.

Lee, M., van Santen, J., Möbius, B., and Olive, J. (1999). Formant Tracking Using Segmental Phonemic Information. In *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*.

Levinson, S. E. (1986). Continuous Variable Duration Hidden Markov Models for Automatic Speech Recognition, *Computer Speech and Language*, 1(1), pages 29-45.

Long, C.J. and Datta, S. (1996). Wavelet Based Feature Extraction for Phoneme Recognition. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 264-267.

Long, C.J. and Datta, S. (1998). Discriminant Wavelet Basis Construction for Speech Recognition. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 1047-1049.

McCandless, S. (1974). An Algorithm for Automatic Formant Extraction Using Linear Prediction Analysis. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume 22, pages 135-141.

Macho, D., Nadeu, C., Hernando, J., and Padrell, J. (1999). Time and Frequency Filtering for Speech Recognition in Real Noise Conditions. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Macho, D. and Nadeu, C. (2001). Comparison of Spectral Derivative Parameters for Robust Speech In *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*.

McClave, J. T., and Sincich, T. (2000). Statistics. 8th ed., Prentice Hall, Upper Saddle River, New Jersey, USA.

McCourt, P., Vaseghi, S. and Harte, N. (1998). Multi-Resolution Cepstral Features for Phoneme Recognition across Speech Sub-Bands. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 557-560.

Mirghafori, N.N. (1999). *A Multi-Band Approach to Automatic Speech Recognition*. Ph.D. thesis, ICSI, Berkeley, California.

Mokbel, C. (1992). *Reconnaissance de la parole dans le bruit : bruitage / debruitage*. Ph.D. thesis, Ecole Nationale Supérieure des Télécommunications, Paris, France.

Mokbel, C., Jouvet, D., and Monné, J. (1996). Deconvolution of Telephone Line Effects for Speech Recognition. *Speech Communication*, volume 19, number 3, pages 185-196.

Mokbel, C. (1999). Personal Communication.

Morgan, N. (1999). Temporal Signal Processing for ASR. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

De Mori, R., Moisa, L., Gemello, R., Mana, F., and Albensano, D. (2001). Augmenting Standard Speech Recognition Features with Energy Gravity Centres. *Computer Speech and Language*, volume 15, pages 341 354.

Morris, A., Hagen, A., Glotin, H. and Bourlard, H. (2001). Multi-stream Adaptive Evidence Combination for Noise Robust ASR. *Speech Communication*.

Nadeu, C. (1999). On the Filter-bank-based Parameterization Front-End for Robust HMM Speech Recognition. In *Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pages 235-238.

Nadeu, C., Hernando, J., and Gorricho, M. (1995). On the Decorrelation of Filter-Bank Energies in Speech Recognition. In *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1381-1384.

Nadeu, C., Macho, D., and Hernando, J. (2001). Time and Frequency Filtering of Filter-bank Energies for Robust HMM Speech Recognition. *Speech Communication*, volume 34, pages 93-114.

Ney, H. (1983). Dynamic Programming Algorithm for Optimal Estimation of Speech Parameter Contours. *IEEE Transactions on Systems, Man, and Cybernetics*, Part A, 13:208-214.

Okawa, S., Bocchieri, E., and Potamianos, A. (1998). Multi-Band Speech Recognition in Noisy Environments. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 641-644.

Olive, J. (1971). Automatic formant tracking in a newton raphson technique. *Journal of the Acoustical Society of America*, 50:661-670.

Olive, J., Greenwood, A., and Coleman, J. (1993). *Acoustics of American English Speech: A dynamic approach*. Springer Verlag, New York.

Ostendorf, M., Digalakis, V., and Kimball, O. (1996). From HMMs to Segment Models: a Unified View of Stochastic Modeling for Speech Recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 4, pages 360--378.

Padmanabhan, M. (2000). Spectral Peak Tracking and its Use In Speech Recognition. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*.

Pearce, D. (1998). Experimental Framework for the Performance Evaluation of Distributed Speech Recognition Front-Ends. Aurora document number AU/120/98.

Peterson, G.E., and Barney, H.L. (1952). Control Methods Used in a Study of the Vowels, *Journal of the Acoustical Society of America*, 24 (2): 175-194.

Rabiner L. and Juang, B. (1993). *Fundamentals of Speech Recognition*. Prentice Hall Signal Processing Series, Englewood Cliffs, NJ.

Rabiner, L., and Schafer, R. (1978). *Digital Processing of Speech Signals*. Prentice Hall Signal Processing Series, Englewood Cliffs, NJ.

Russel, M.J. and Moore, R.K. (1985). Explicit Modelling of State Occupancy in Hidden Markov Models for Automatic Speech Recognition. *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 5-8.

Samaria, F. (1994). *Face Recognition Using Hidden Markov Models*. Ph.D. thesis, Engineering Department, Cambridge University.

Schafer, R.W. and Rabiner, L.R. (1970). System for Automatic Formant Analysis of Voiced Speech. *Journal of the Acoustical Society of America*, 57(634-648).

Schmid, P., and Barnard, E. (1995). Robust, N-best Formant Tracking. In *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 737-740.

Schmid, P. (1996). *Explicit N-best Formant Features for Segment-Based Speech Recognition*. Ph.D. thesis, Department of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, Portland.

Smyth, P., Heckerman, D., and Jordan, M. (1997). Probabilistic Independence Networks for Hidden Markov Probability Models. In *Neural Computation*, volume 9, number 2, pages 227-269.

SPHEAR TMR NETWORK project homepage, http://www.dcs.shef.ac.uk/~pdg/sphear/sphear.htm.

Stephenson, T. A. (2003). Speech Recognition with auxiliary information. Ph.D. thesis, Ecole Polytechnique Federale de Lausanne, Switzerland.

Stuttle, M. and Gales, M.J.F. (2001). A Mixture of Gaussians Front End for Speech Recognition. In *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*.

Talkin, D. (1987). Speech Formant Trajectory Estimation Using Dynamic Programming with Modulated Transition Costs. AT&T Internal Memo MH 11222 2924 2D-410, AT&T.

Varga, A.P. and Moore, R.K. (1990). Hidden Markov Model Decomposition of Speech and Noise. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 845-848.

Varga, A.P. and Moore, R.K. (1991). Simultaneous Recognition of Concurrent Speech Signals using Hidden Markov Model Decomposition. In *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1175-1178.

Varga, A., Steeneken, H.J.M., Tomlinson, M. & Jones, D. (1992). The Noisex-92 Study on the Effect of Additive Noise on Automatic Speech Recognition. Technical Report, DRA Speech Research Unit.

Viterbi, A.J. (1967). Error bounds for convolution codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, 13:260-269.

Wang, X. (1997). *Incorporating Knowledge on Segmental Duration in HMM-Based Continuous Speech Recognition*. Ph.D. thesis, University of Amsterdam, The Netherlands.

Wassner, H. and Chollet, G. (1996). New Cepstral Representation Using Wavelet Analysis And Spectral Transformation For Robust Speech Recognition. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*.

Weber, K. (2000). Multiple Timescale Feature Combination towards Robust Speech Recognition. In *KONVENS 2000 / Sprachkommunikation*, Ilmenau, Germany, pages 295-299.

Weber, K., Bengio, S., and Bourlard, H. (2000). HMM2- A Novel Approach to HMM Emission Probability Estimation. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume III, pages 147-150.

Weber, K., Bengio, S., and Bourlard, H. (2001a). HMM2- Extraction of Formant Structures and their Use for Robust ASR. In *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 607-610.

Weber, K., Bengio, S., and Bourlard, H. (2001b). A Pragmatic View of the Application of HMM2 for ASR. Technical Report IDIAP-RR 01-23. ftp://ftp.idiap.ch/pub/reports/2001/rr01-23.ps.gz.

Weber, K., Bengio, S., and Bourlard, H. (2001c). Speech Recognition Using Advanced HMM2 Features. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Weber, K., Bengio, S., and Bourlard, H. (2002). Increasing Speech Recognition Noise Robustness with HMM2. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 929-932, Orlando, Florida, USA.

Weber, K., de Wet, F., Cranen, B., Boves, L., Bengio, S., and Bourlard, H. (2002). Evaluation of Formant-Like Features for ASR. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 2101-2104, Denver, CO, USA.

Weber, K., Ikbal, S., Bengio, S., and Bourlard, H. (2003). Robust Speech Recognition and Feature Extraction Using HMM2. In *Computer Speech and Language*, volume 17/2-3, pages 195-211.

Wellekens, C.J. (1987). Explicit Time Correlation in Hidden Markov Models for Speech Recognition. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 384-386.

Welling, L. and Ney, H. (1998). Formant Estimation for Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, volume 6, number 1, pages 36-48.

Werner, S. and Rigoll, G. (2001). Pseudo 2-dimensional Hidden Markov Models in Speech Recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

De Wet, F., Cranen, B., de Veth, J., and Boves, L. (2000). Comparing Acoustic Features for Robust ASR in Fixed and Cellular Network Applications. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 1415-1418.

De Wet, F., Weber, K., Cranen, B., Boves, L., Bengio, S., and Bourlard, H. (2002). Evaluation of Formant-Like Features for Automatic Speech Recognition. Technical Report IDIAP-RR 03-08, submitted for publication in the *Journal of the Acoustical Society of America (JASA)*.

Willems, L.F. (1986). Robust Formant Analysis. IPO Annual report 21, Eindhoven, The Netherlands, pages 34-40.

Wu, S., Kingsbury, B., Morgan, N., and Greenberg, S. (1998). Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 721-724.

Young, S., Odell, J., Ollason, D., Valtchev, V. and Woodland, P. (1995). *The HTK Book*. Cambridge University, UK.

Zolfaghari, P. and Robinson, A.J. (1996). Formant Analysis using Mixtures of Gaussians. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*.

Zweig, G. G. (1998). *Speech Recognition with Dynamic Bayesian Networks*. Ph.D. thesis, University of California, Berkeley.

Curriculum Vitae

Name:	Katrin Weber, née Keller
Date of Birth:	11 September 1969
Place of Birth:	Sangerhausen, Germany

CURRENT POSITION

Jan. 98 to date:	Research assistant / Ph.D. student in the Speech Group at IDIAP- Dalle Molle Institute of Perceptual Artificial Intelligence, Martigny, Switzerland.

EDUCATION

1989 - 1997	Study of Computer Science at the Technical University of Ilmenau, Faculty of Computer Science and Automatic Systems. Diploma in Computer Science.
1992 - 1993	Special Course: BSc 3rd year course in Computer Science and special project at the University of Exeter, Department of Computer Science, Great Britain.
1986 - 1989	Abitur (German equivalent of A-levels), in combination with vocational training, at the former "Robotron Elektronik" Zella-Mehlis. Qualification as data processing engineer.
1976 - 1986	Polytechnische Oberschule (10-class comprehensive school).

PROFESSIONAL EXPERIENCE

Aug. 95 - Feb. 96	Research in the field of Neural Networks at IDIAP- Dalle Molle Institute of Perceptual Artificial Intelligence, Martigny, Switzerland.
1992, 1994	Tutorial assistant at the Faculty of Computer Science and Automatic Systems, Technical University of Ilmenau.
Nov. 93 - Feb. 94	ISDN device driver programming at the IBM European Networking Centre, Heidelberg.
Aug. - Sep. 1991	Database design and programming at Sequent Computer Systems GmbH, München.

MISCELLANEOUS SKILLS AND INTERESTS

Languages: German (mother tongue), fluent English and French, adequate Russian; Car driver (clean licence); Photography, Traveling, Mountains.

PUBLICATIONS

[1] Keller (now Weber), K., Ben-Yacoub, S., and Mokbel, C. (1999). Combining Wavelet-domain Hidden Markov Trees with Hidden Markov Models. IDIAP-RR 99-14.

[2] Bengio, S., Bourlard, H., and Weber, K. (2000). An EM Algorithm for HMMs with Emission Distributions Represented by HMMs, IDIAP-RR 00-11.

[3] Weber, K. (2000). Multiple Timescale Feature Combination towards Robust Speech Recognition. In *KONVENS 2000 / Sprachkommunikation*, pages 295-299, Ilmenau, Germany.

[4] Weber, K., Bengio, S., and Bourlard, H. (2000). HMM2- A Novel Approach to HMM Emission Probability Estimation. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume III, pages 147-150, Beijing, China.

[5] Bourlard, H., Bengio, S., and Weber, K. (2001). New Approaches Towards Robust and Adaptive Speech Recognition. In *Advances in Neural Information Processing Systems 13,* MIT Press.

[6] Ikbal, S., Bourlard, H., Bengio, S., and Weber, K. (2001). IDIAP HMM/HMM2 System: Theoretical Basis and Software Specifications. IDIAP-RR 01-27.

[7] Weber, K., Bengio, S., and Bourlard, H. (2001a). HMM2- Extraction of Formant Features and their Use for Robust ASR. In *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 607-610, Aalborg, Denmark.

[8] Weber, K., Bengio, S., and Bourlard, H. (2001b). A Pragmatic View of the Application of HMM2 for ASR. IDIAP-RR 01-23.

[9] Weber, K., Bengio, S., and Bourlard, H. (2001c). Speech Recognition Using Advanced HMM2 Features. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Madonna di Campiglio, Italy.

[10] Weber, K., Bengio, S., and Bourlard, H. (2002). Increasing Speech Recognition Noise Robustness with HMM2. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 929-932, Orlando, Florida, USA.

[11] Ikbal, S., Weber, K., and Bourlard, H. (2002). Speaker Normalization using HMM2. In *Proceedings of the International IEEE Workshop on Neural Networks for Signal Processing (NNSP 2002)*, pages 647-656, Martigny, Switzerland.

[12] Weber, K., de Wet, F., Cranen, B., Boves, L., Bengio, S., and Bourlard, H. (2002). Evaluation of Formant-Like Features for ASR. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 2101-2104, Denver, CO, USA.

[13] Bourlard, H., Bengio, S., and Weber, K. (2003). Towards Robust and Adaptive Speech Recognition Models. To appear in Ostendorf, M., Khudanpur, S., and Rosenfeld, R., editors, *Mathematical Foundations of Speech Processing and Recognition*, Institute for Mathematics and its Applications (IMA) Series. Springer-Verlag.

[14] Weber, K., Ikbal, S., Bengio, S., and Bourlard, H. (2003). Robust Speech Recognition and Feature Extraction Using HMM2. In *Computer Speech and Language*, volume 17/2-3, pages 195-211.

[15] de Wet, F., Weber, K., Cranen, B., Boves, L., Bengio, S., and Bourlard, H. (2003). Evaluation of Formant-Like Features for Automatic Speech Recognition. IDIAP-RR 03-08, submitted for publication in the *Journal of the Acoustical Society of America (JASA)*.